Developing a content and knowledge-based journal recommender system comparing distinct subject domains

Dissertation
zur Erlangung des akademischen Grades

Doctor philosophiae
(Dr. phil.)

im Fach Bibliotheks- und Informationswissenschaft
eingereicht

an der Philosophischen Fakultät I
der Humboldt-Universität zu Berlin

von Manjula Wijewickrema

Die Präsidentin der Humboldt-Universität zu Berlin
Professor Dr.-Ing. Dr. Sabine Kunst

Die Dekanin der Philosophischen Fakultät I
Professor Dr. Gabriele Metzler

Gutachter/Gutachterinnen
1. Professor Vivien Petras, PhD
2. Professor Naomal Dias, PhD

Datum der Einreichung: 06.05.2019
Datum der Disputation: 24.06.2019

Abstract

With the rapidly growing number of journal outlets being produced every year, authors need assistance in selecting the most appropriate journal outlet for submitting manuscripts. The task of finding appropriate journals cannot be accomplished manually due to a number of limitations of the approach. This becomes more complicated as the selection criteria may change from one discipline to another. Therefore, to address this issue, the current research develops a journal recommender system with two components: the first component compares the content similarities between a manuscript and the existing journal articles in a corpus. This represents the content-based recommender component of the system. In addition, the system includes a knowledge-based recommender component to consider authors' publication requirements based on 15 journal selection factors. The new system makes recommendations from the open access journals indexed in the directory of open access journals for two distinct subject domains, namely medicine and social sciences.

The study initially compared 16 journal factors that could influence journal choices. A web-based survey was conducted to collect information from authors who have recent publications in open access journals. According to the results, authors of both subject areas acknowledged 'peer-review' status as the most important factor, while giving least attention to the 'number of annual subscribers' of the journal. The results were used further to determine the importance of each factor from authors point of view. These factor weights were expected to use for implementing the knowledge-based recommender component. Next, appropriateness of five algorithms was examined to select the best one to implement the content-based recommender component. Overall results revealed that the BM25 similarity outperforms the other four algorithms considered in the study. The unigram language measure showed the lowest performance. A knowledge-based recommender component was developed to merge with the content-based recommender component. This component arranges the order of journals suggested by the content-based component based on au-

thor's criteria of journal selection. The Gower's measure was implemented to determine the similarity between author's selection criteria and journal's available criteria. Then, a second author survey was conducted to collect information to configure the hybrid recommender system that merged content and knowledge-based components. The survey asked the authors whether and to what extent they considered the given 15 journal factors when selecting an appropriate journal for one of their recently published articles. A third author survey allowed respondents of the second author survey to rank the appropriateness of journals suggested by the hybrid recommender system from their point of view. The results indicated that the authors from medicine and social sciences agree with the recommender's suggestions by 66.2% and 58.8% respectively. Moreover, 35.5% of medicine and 40.4% of social sciences authors were suggested more appropriate journal(s) than the journal they already published in. Average performance of the system demonstrated 15% and 18% performance loss in medicine and social sciences respectively against the same suggestions after arranging according to the most appropriate order. Numbers were reported as 22.4% and 28.4% of loss in medicine and social sciences respectively when the average performance was compared with a system that retrieves appropriate suggestions for all 10 topmost results according to the most appropriate order. Although the hybrid recommender demonstrated a slight advancement of performance than the content-based component, the improvement was not statistically significant.

The outcome of the research is useful for many other parties, in addition to authors. Among the others, journal editors, publishers, policy developers for academic institutions, and system developers can benefit directly with the new journal recommender system. To the best of our knowledge, the current research is the first one to develop a journal recommender system combining a content-based component with a knowledge-based component associated with 15 factors, using a similarity measure. Moreover, no previous work has focused on investigating potential influence of numerous properties of distinct subject corpora for the performance of journal recommender systems.

Zusammenfassung

Bei der von Jahr zu Jahr schnell wachsenden Anzahl von publizierten Journals benötigen Autoren Hilfe bei der Wahl des Journals, das zum Einreichen eines Manuskripts am besten geeignet ist. Die Aufgabe, ein passendes Journal zu finden, ist auf Grund von verschiedenen Einschränkungen nicht von Hand zu erledigen. Da sich die Auswahlkriterien je nach Disziplin unterscheiden können, wird die Sache noch komplizierter. Um also diese Problematik zu behandeln, entwickelt die aktuelle Untersuchung ein Journal-Empfehlungssystem, das – in einer Komponente – die inhaltlichen Ähnlichkeiten zwischen einem Manuskript und den existierenden Zeitschriftenartikeln in einem Korpus vergleicht. Das stellt die inhaltsbasierte Empfehlungskomponente des Systems dar. Zusätzlich beinhaltet das System eine wissensbasierte Empfehlungskomponente, um die Anforderungen des Autors bezüglich der Veröffentlichung auf Basis von 15 Journal-Auswahlkriterien zu berücksichtigen. Das neue System gibt Empfehlungen aus den im Directory of Open Access Journals (DOAJ) indizierten Journals für zwei verschiedene Themengebiete: Medizin und Sozialwissenschaften.

Ursprünglich verglich die Studie 16 Faktoren, die die Auswahl eines Journals beeinflussen. Eine webbasierte Befragung sammelte Informationen von Autoren, die vor kurzem einen Artikel in einem Open-Access-Journal veröffentlicht haben. Den Ergebnissen zufolge wird die Begutachtung durch Kollegen ('peer-review') von Autoren in beiden Disziplinen als wichtigster Faktor angesehen, während sie die Anzahl von jährlichen Abonnenten des Journals am wenigsten beachten. Die Befragungsergebnisse wurden weiterverwendet, um die Wichtigkeit eines jeden Faktors aus Sicht der Autoren zu bestimmen. Es wurde erwartet, dass die Gewichtung der Faktoren für die Implementierung der wissensbasierten Empfehlungskomponente eingesetzt wird. Als nächstes

wurde die Eignung von fünf Algorithmen zur Ähnlichkeitsbestimmung untersucht, um den davon am besten geeigneten für die Implementierung der inhaltsbasierten Empfehlungskomponente zu verwenden. Die Ergebnisse zeigen, dass das BM25 Ähnlichkeitsmaß die anderen vier in der Studie betrachteten Algorithmen übertrifft. Das Unigram-Maß bot die niedrigste Leistung. Eine wissensbasierte Empfehlungskomponente wurde zur Zusammenlegung mit der inhaltsbasierten Empfehlungskomponente entwickelt. Diese Komponente ordnet die von der inhaltsbasierten Komponente vorgeschlagenen Journals auf Basis der Journal-Auswahlkriterien des Autors. Das Distanzmaß nach Gower wurde implementiert, um die Ähnlichkeit zwischen den Auswahlkriterien des Autors und den verfügbaren Kriterien des Journals zu bestimmen. Dann wurde eine zweite Befragung durchgeführt, um Informationen für die Konfiguration des hybriden Empfehlungssystems, das inhaltliche und wissensbasierte Komponenten zusammenbringt, zu sammeln. Autoren wurden gefragt, ob und inwieweit sie die gegebenen 15 Journal-Auswahlkriterien bei der Selektion eines geeigneten Journals für einen kürzlich veröffentlichten Zeitschriftartikel berücksichtigten. Im Rahmen einer dritten Befragung erstellten die Teilnehmer der zweiten Befragung eine Rangliste der vom hybriden Empfehlungssystem vorgeschlagenen Journals bezüglich der Eignung aus ihrer Sicht. Die Ergebnisse zeigen, dass die Autoren aus den Themengebieten Medizin und Sozialwissenschaften mit den Empfehlungen des Systems zu 66,2% bzw. 58,8% einverstanden waren. Darüber hinaus wurde 35,5% der Autoren aus dem Bereich Medizin und 40,4% der Autoren aus den Sozialwissenschaften ein oder mehrere Journal(s) vorgeschlagen, das bzw. die für die Publikation besser geeignet war(en) als das Journal, in dem sie den Artikel veröffentlich hatten. Die durchschnittliche Leistung des Systems zeigte eine Abnahme von 15% in Medizin bzw. 18% in Sozialwissenschaften verglichen mit den gleichen Empfehlungen bei einer optimalen Sortierung. Leistungsverluste von 22,4% im

Fach Medizin und 28,4% in den Sozialwissenschaften ergaben sich, wenn die durchschnittliche Leistung mit einem System verglichen wurde, das geeignete Empfehlungen für die 10 besten Resultate in der optimalen Reihenfolge sortiert abruft. Die vom Hybrid-Modell Empfehlungen zeigen zwar eine etwas bessere Leistung als die inhaltsbasierte Komponente, die Verbesserung war aber nicht statistisch signifikant.

Die Ergebnisse dieser Forschung sind nicht nur für Autoren nützlich, auch Herausgeber, Verleger, Entwickler von Richtlinien für wissenschaftliche Einrichtungen, Systementwickler u. a. können direkt von dem neuen Journal-Empfehlungssystem profitieren. Nach bestem Wissen ist diese Dissertation die erste, die ein Journal-Empfehlungssystem mit einer Kombination aus einer inhaltsbasierten Komponente und einer wissensbasierten Komponente entwickelt hat, in dem 15 Faktoren und einen Algorithmus zur Ähnlichkeitsbestimmung eingesetzt werden. Außerdem fokussierte keine vorherige Arbeit die Untersuchung des potentiellen Einflusses von zahlreichen Eigenschaften verschiedener Korpora unterschiedlicher Fächer auf die Leistung von Journal-Empfehlungssystemen.

Acknowledgements

Foremost, I would like to express the deepest appreciation to my principal advisor Professor Vivien Petras, Berlin School of Library and Information Science, Humboldt University of Berlin, Germany, for the patient guidance, constant encouragement, and immense knowledge contributed throughout my time as her student. This research would not have been materialized without her incisive observations and intellectual directions. I have been very fortunate to have an advisor who cared a lot about this work and responded to my queries so promptly. While steering me through the correct scientific method, her excellent skills of scientific communication supported me to sharpen my academic writing too. She has taught me more than I could ever give her credit in this little space. I owe a deep of gratitude to the co-advisor Professor Naomal Dias, Department of Computer Systems Engineering, University of Kelaniya, Sri Lanka, for his invaluable inputs at crucial points of the research. Short scientific discussions had with him helped me to continue this study on the correct track.

I must express my sincere gratitude to Professor Elke Greifeneder and few other academics at the Berlin School of Library and Information Science, Humboldt University of Berlin, Germany, and my colleagues at the Sabaragamuwa University of Sri Lanka for participating in pre-tests and sharing their important ideas to improve the surveys. All anonymous participants of the surveys including journal editors/editorial staff are highly appreciated for being part of my dissertation work. Further, I humbly extend my thanks to all the members of staff at the Berlin School of Library and Information Science for providing me office space with needed resources to conduct the research.

This work would not have been possible without the financial support of the National Centre for Advanced Studies in Humanities and Social Sciences, Colombo, Sri Lanka. I am especially indebted to my employer, Sabaragamuwa University of Sri Lanka for granting me study leave for reading for doctoral

studies overseas.

Last but not the least, I would like to thank my parents, whose love and guidance are with me in whatever I pursue. Most importantly, I wish to thank my loving and supportive wife, Upeksha for her views, suggestions, motivation, and support for formatting the dissertation. To conclude, I remind my little son, Imeth, who provided unending inspiration to succeed difficult times of this long run.

Contents

List of Tables

List of Figures

List of Abbreviations

A&I	Abstracting and Indexing
APC	Article Processing Charge
API	Application Programming Interface
C&K	Content-based and Knowledge-based
DCG	Discounted Cumulative Gain
DOAJ	Directory of Open Access Journals
ERR	Expected Reciprocal Rank
IDCG	Ideal Discounted Cumulative Gain
IF	Impact Factor
IR	Information Retrieval
JCR	Journal Citation Reports
JDBC	Java Database Connectivity
LCC	Library of Congress Classification
MAgP	Mean Average generalized Precision
MAP	Mean Average Precision
MNB	Multinomial Naïve Bayes

NDCG	Normalized Discounted Cumulative Gain
OA	Open Access
ODBC	Open Database Connectivity
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristics curve
SJR	SCImago Journal Rank
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
VSM	Vector Space Model
WoS	Web of Science

Chapter 1

Introduction

"Publish, then no need to perish"

– Author

The growth of research in various disciplines leads people to document the research process properly, making them available for others to be informed about the innovation, and impels others to study more about the innovation, if it is interesting for them. Academic journals can be considered one of the most accepted approaches in existence to accomplish the publication task of a manuscript authored by a researcher to appraise their achievement. They also act as carriers which bring readers to the innovation and allow them to explore the subject matter further. The nature of the publishing process could vary from simple to very complex based on the purpose of publication, significance and validity of contents, intended audience, medium of publication, behavior of authors, publishers, editors, reviewers, funders, plus many other factors. Therefore, it is not reasonable to expect a smooth progression throughout the publishing process since the negative impacts of influential factors are highly likely to daunt the process at some point. This dissertation challenges one such vital issue that serves as an impediment in the publishing arena.

1.1 Publishing

The broad idea behind publishing as an industry is the dissemination of information, making it more widely accessible and available for public consumption. It took centuries after the innovation of writing, for the writing of public intellectuals to become available to the masses after the crucial intervention of the printing machine - a leap in mankind's achievements which began progressing much more rapidly thereafter. At present, publishing has become a part of people's day-to-day activities, spread over multiple fields. News publishing via print and electronic media is perhaps the most frequent application of publishing due to massive demand from the public. Book publishing is also a well-known format of publishing of information. Possibility of including lengthy information could be the greatest advantage of this publication format. Books are published in both print and electronic forms currently, although readers still prefer printed books over e-books, citing comfort. Periodical publishing includes information of subject matter in detail. Usually, information included in periodicals may attain higher reliability compared to information in newspapers. A regular interval between two issues is one of the noticeable characteristics of this publishing system. Directory publishing includes a short snippet of information to readers that usually directs to a source with comprehensive information. For example, *Yellow Pages* is a renowned contact directory that includes telephone numbers and mailing addresses of companies and individuals. Similar to periodicals, most directories are issued under regular time periods. Standards and patents communicate specific information to readers in brief. Information included in them is often used by researchers and technicians, but may not be as important for the general community. In addition to these, publishing in ordinary websites and established social media is becoming quite popular at present.

Academic publishing can appear in both periodical and book formats in addition to theses and proceedings at conferences or workshops. Short academic articles are published occasionally in newspapers and frequently in ordinary websites though they do not receive higher recognition due to lack of authentication (Ortiz et al.,

2005; Seife, 2014). Academic articles included in standard media are concentrated in specific disciplines, and some are reviewed by peers before publishing. Introduction of business models to academic publishing provides numerous options for authors and readers to publish or subscribe to source contents. To illustrate, Open Access (OA) and fee-based access models are available for readers to reach published interesting articles, while authors are allowed to submit manuscripts with an Article Processing Charge (APC) or without an author charge. At present, the area of academic publishing has widely been expanded parallel to providing rapid publication opportunities in digital media (Ronte, 2001). The topic covered by this dissertation describes an underlying problem in academic publishing, while specially emphasizing its influence with regards to scholarly journals.

Obviously, the primary goal of academic publishing directs towards dissemination of new knowledge to others. Sharing new knowledge is not only crucial for continuous progress of an academic discipline, but also important for career development of a researcher.

Publishing in reputed journals is extremely important for scholars to achieve their professional goals. Among others, promotions in careers basically depend on publication factor in most academic institutions (Langston, 1996). Not only promotions from junior level to senior level, but also promotions to next grade of the same seniority level can be achieved with sufficient publications in scholarly journals. This factor can be also considered as an indicator to measure a candidate's career development parallel to promotions obtained.

A list of relevant publications in the field of a new occupation would give an additional shine to someone's curriculum vitae. Properly compiled list of publications can easily be used to reflect candidate's exposure in the field of interest. A new employer would also be benefited by the publication list as it can be used to filter the best among many competitive candidates.

Publishing in high impact journals indexed or abstracted by leading databases is en-

couraged by institutional rewards. Receiving financial assistance to cover submission charges and APC of accepted manuscripts in reputed journals is an added advantage of submitting to high impact journals. In addition to a good research proposal, prior publications in the relevant area can make a significant impact in securing funds for new research projects.

Recognition among peers is important to build new academic networks of researchers in a considered field. The topics covered by a list of publications are helpful to identify the researchers working in a particular area. This ultimately leads to better communication and effective dissemination of knowledge among them. Publishers could utilize author information in already published articles to appoint appropriate editors for their journals. Editors often use existing publications to identify potential reviewers for manuscripts. Conference organizers also follow this methodology to invite relevant researchers to present new findings at their events. Recognition among peers is important not only for journals and conferences, but also to showcase one's name as an eminent researcher. This actively encourages scholars to carry out diligent research in their areas of expertise.

Some of the research based postgraduate degree programmes evaluate publications in addition to the final dissertation of the study. However, an extremely selective criterion of a journal in which the articles must be published could be an enormous challenge for most postgraduate students with inadequate experience in publishing. Accomplishment of this goal indicates the significance, validity, and novelty of the work from the view point of external reviewers who do not engage with the study at all. Furthermore, already published articles under an ongoing research topic of the degree programme may reduce the higher weight usually assigned to the final dissertation.

1.2 The Problem

There is no reason to believe that the present scholars are not enthusiastic to make publications, but we must not underestimate the numerous difficulties they face throughout the publication process. The rapid increase of the number of journals over the last eight years (Cope and Phillips, 2014; Larsen and von Ins, 2010; Solomon et al., 2013), have opened a wide opportunity for the authors to publish their works without much difficulty. Growth of online journals aids much in boosting this opportunity at present. Technological advancements like online submission facilities smoothen the process even further for current day authors. Moreover, the availability of journals with free of author charges encourages researchers to write and publish their manuscripts without stressing about financial constraints. Thus, the authors in the modern research culture are motivated, and provided numerous journal options to publish their exhaustive efforts. Nevertheless, it is to be noted, the same authors could be confused by the abundance of publication options, while selecting the most appropriate option for submitting their research. This situation occurs due to a number of reasons such as the author's lack of knowledge about the subject coverage of the target journal, inability to identify the rank, reputation and standard of the journal, underestimation of author's own research and overestimation of the journal (and vice versa), and unawareness of the timelines of publication of the journal. Therefore, the authors have to devote an extra effort to evaluate the related journals one by one. Even so, these evaluations may not yield practical results owing to practical difficulties such as the impossibility of considering all related journals for the evaluation, difficulty of finding a standard evaluation criterion for the journal selection, and the lengthy time commitments required for the evaluation of each related journal manually. Furthermore, the mere selection of journal outlets leads to a number of problems. The immediate rejection of a submission due to out of scope issues discourages authors and consequently leads to a decrease in the productivity of the researcher. An outdated publication, after going through several unsuccessful attempts at inappropriate journals, would have less impact than originally capable

of, on the research community. Even if the research is relevant, society will not be able to get its benefit unless it appears on time. In addition to the number of attempts, another reason for outdated articles is the relatively long time span that journals take to complete the review process and to make a decision about the acceptance of a manuscript. For instance, if a manuscript is rejected after the final revision of its first submission, the author has to initiate the submission process yet again for another journal outlet. Finally, even if the manuscript is accepted by a second journal, the waiting time after the first submission could lead to publishing an obsolete work. Negative outcomes of selecting an inappropriate publishing outlet for submitting the manuscript can lead to frustrating authors in different ways. Manuscript rejections, publication delays, and obsolete publications, all could effect the author's job interviews, promotions at their current occupation, institutional rewards and many further achievements. Hence, a study on implementing a system to identify the most suitable journal outlet with the highest possibility of publishing an article, which is also compatible with the author's needs will definitely encourage the authors to write and communicate more. The proposed research is focused on this critical practical problem, which is as yet unsolved.

This prominent question of scholarly communication cannot be addressed by the available manual methods alone. On the one hand, selecting the most appropriate journal outlet from thousands of possible options is a labor intensive work. On the other hand, the process may involve considerable amounts of computational work since the comparison between the manuscript and the potential journals must use a scoring methodology to prioritize the options. Thus, there is an inevitable need of amalgamation of this research area with a possible application domain in computer science. Incorporating text mining methodologies can be seen as a possible solution for the problem since processing and analyzing data in large-scale are being deployed vastly for mining and extracting different texts. Experimental outcomes of text classification have even been applied in some real world situations like plagiarism detection, astroinformatics, astrostatistics, medicine, criminology, and spam filtering. Moreover, a number of studies have already been done on recommending journal

outlets for given manuscripts. Some of them went even beyond the conceptual level and developed tools like Elsevier Journal Finder[1], Journal/Author Name Estimator (JANE)[2], Springer Journal Advisor[3], and eTBLAST[4] to select the best journal. Nevertheless, these systems have certain drawbacks which prevent optimal results somehow. These systems and their drawbacks are discussed in sections 2.2 and 2.6.

Therefore, in this study, a new approach for text classification, for selecting the most appropriate journal outlet for manuscript submission based on large text datasets is developed. The new journal recommender system intends to eradicate the drawbacks which were identified in the previous solutions. Instead of focusing on a single factor, the improvements are achieved by addressing the problem in a number of aspects such as establishing a more appropriate similarity measure to compare the manuscript and journal outlets, searching/retrieving journals from multiple subject domains with distinct properties, and utilizing a sufficiently large portion of the text to determine the similarities.

1.3 Medicine, social sciences, and open access journals

1.3.1 Two subject domains: Medicine and social sciences

The current study selects two widely different subject domains to implement the new journal recommender system. More than two subject domains are not considered due to the time constraints. The main objective of selecting medicine and social sciences domains is based on their clearly visible distinctions associated with the publication process. These distinctions can be categorized under two major topics.

[1] http://journalfinder.elsevier.com/

[2] http://www.biosemantics.org/jane/

[3] http://www.springer.com/gp/authors-editors/journal-author/journal-author-helpdesk/preparation/1276

[4] http://etest.vbi.vt.edu/etblast3/

1. Distinctions of the text of the journal articles: vocabulary used, length of the articles, frequency of similar terms, and so on.

2. Distinctions of the publication culture: qualitative and quantitative publication factors (see section 3.2) considered by the authors and the publication outlets of the two subject domains. For example, qualitative aspects such as quality of the contents and significance of the study may be considered more in some subject domains than in others.

Contents of the articles in medicine may include more technical terms than in an article from the domain of social sciences. Technically enriched vocabularies can also be found from the domains such as mathematics, information technology, engineering, and so on. This is usually a common characteristic of the subject domains in natural sciences. In contrast, subject domains such as social sciences and humanities may include fewer amounts of technical terms, but with higher diversity of terms from the vocabulary used in general. In addition, the frequency of the terms used in the articles may depend on the subject domain. Therefore, it is highly likely to differentiate the behavior of a journal recommender system based on these distinct textual properties of the subject domain they work.

Further, the nature of the publication process between medicine and social sciences may vary from one subject domain to another. Significantly different values of the quantitative factors that involved with the publication process can be considered as a reflection of these differences. For instance, the behavior of journal Impact Factor (IF) is highly depending on the subject domain of journals. This is a common characteristic of the citation based indices that used to rank journals (González-Betancor and Dorta-González, 2017). Moreover, factors such as publication frequency and the number of articles published per issue could change widely based on the rate, a subject domain is updated. To illustrate, a journal in a rapidly changing subject domain such as medicine or information technology could produce more issues and more articles per year than a journal belongs to history. Therefore, the current study considers the impact of these widely different publication factors in distinct subject

domains for implementing the recommender system in medicine and social sciences separately.

Apart from the requirement to study the behavior of the proposed recommender system in two radically distinct subject domains, the importance of the two subject domains attained in the area of publishing is considered to include medicine and social sciences in the current study. On the one hand, research and publications in medicine are important since the outcomes of this domain are the potential implications of people's health risks and the measures, which can minimize the risks. On the other hand, research in the domain of social sciences contributes to the progress of the society based on numerous aspects including economics, politics, education, culture, and so on. Therefore, the sustainable development of a society is likely to depend on the continuous growth of the social sciences subject domain with the support of other subjects too. Usually, the two subject domains – medicine and social sciences contribute a large amount of scholarly literature to the progress of the society. A lot of publications can be found from medicine (Dixon-Woods and Tarrant, 2009; Nelson, 2009; Shavers-Hornaday et al., 1997) and social sciences (Arendell and Reinharz, 1995; Groneberg, 2018; Landry et al., 2001), because of the higher attention these subjects receive. Moreover, due to the same reason, the scholarly publishers are enthusiastic to publish a considerable amount of journals including these two subject domains. For example, the Elsevier database included 26% health sciences literature by August 2017, while 31% of database literature was covered by social sciences (Elsevier, 2017). Therefore, considering all these important characteristics of medicine and the social sciences, the novel journal recommender utilizes them as the subject domains for comparison.

1.3.2 Open access journals

Selecting OA journals for the current study is based on three major reasons.

1. Significance of OA publishing.

2. Influence of the different nature of publication factors for selecting OA journals.

3. Feasibility of using OA journals for the current study.

Free of charge access to the contents of journals could be the major advantage of OA publications compared to the non-OA publications from the reader's point of view. The growth of OA journals has been increasing dramatically since 1990s due to their higher popularity among scholars (Solomon et al., 2013). Author's willingness to receive relatively higher number of citations for published articles in OA journals, while readers have more opportunity to access them could be the key reason for their increasing popularity. Further, higher IF values attained by some OA mega journals could be among the reasons for attracting authors towards OA journals. Most of the existing journal recommender systems are based on the prominent commercial journal databases such as Science direct, Springer, IEEE, and so on. Although some of these recommenders allow making recommendations from small collections of OA journals, they do not search across large OA databases like Directory of Open Access Journals (DOAJ). As a result, there is a higher possibility of missing the most appropriate journal for an author's submitted article. Thus, the novel journal recommender targets to bridge this existing gap in journal recommender research.

OA or non-OA status of a journal could also influence on the journal selection factors such as circulation and IF. For example, since the all OA journals are available online and free of charge, more circulation can be expected from them than the non-OA journals. Therefore, the difference of this factor could be considerably higher between OA and non-OA journals. As a result, searching appropriate journals in both OA and non-OA databases simultaneously for the same factor values would lead to generate less precise recommendations. However, the proposed method avoids this drawback since focus only on OA journals for recommendations.

Finally, the feasibility of using OA journals is considered by the current study to include only OA journals. The two subject corpora discussed in section 3.4.2 are based on a large collection of full-text articles. Obviously, full-text articles include

more information than short texts like abstracts. Therefore, we decided to include full-text articles in the training corpora as it could lead the recommender system to generate more accurate results with more information in the texts. However, collecting a large number of recently published full-text articles from a number of commercial journal databases is challenging due to the access limitations to them. Moreover, using full-text articles collected from the commercial journal databases would require proper permission from individual publishers of the journals due to their copyright laws. Thus, the current study targets to include OA journals for the novel recommender system considering the feasibility of accessing and extracting the full-texts with their flexible copyright laws.

1.4 Available methods to assist

One can find a number of methods which could assist an author in selecting an appropriate journal outlet for submitting a manuscript. Some of the methodologies have been established decades earlier, while others are relatively new. Almost all these methodologies include both pros and cons that deter the author from selecting the most appropriate venture.

Experienced researchers who have been publishing for several years could assist a layman when choosing a proper publication outlet for a given manuscript. This approach is also known as colleagues' or peers recommendation. Extensive exposure to the publication process in their disciplines can be considered as an adequate qualification to recommend an appropriate journal to others. Nevertheless, this is no longer a valid argument since rarely could anyone have an informed idea about all the available publication outlets even in his or her field of expertise. The rapid growth of journals increases the possibility of missing an appropriate outlet if someone attempts to select a journal based on his or her experience. Introduction of new publication outlets is highly likely to increase the possibility of selecting a more suitable journal, which was not there in the days prior. Unfortunately, authors cannot update their

knowledge-base about publications at the same rate that publication venues are growing. As a result, it is inevitable that even the experienced authors will miss a more appropriate journal. Consulting librarians is also similar to this approach, but may include more pitfalls as they might not have the knowledge about publications in very specific areas.

Selecting a journal from recommended journal lists compiled by the author's institution is especially useful for receiving institutional rewards and promotions of job tenure. Selecting an appropriate journal from comparatively less number of options keeps authors more stress-free since the authors have to make only a few comparisons. Moreover, most institutions encourage the employees to submit their manuscripts to listed journals by providing submission and APC if the manuscripts are accepted for publication. However, a number of disadvantages innate to this approach prevent the authors from selecting this option more frequently. Limited number of journals in the list is a severe drawback of this approach. The lesser number of selections means that the author may be deprived of the most suitable journal, even though it is easy to compare a short list. Furthermore, the personal preferences of the members of journal selection committee will result in a biased list of journals, which may not be applicable for an average author. Visibility of the journals included in an institutional recommended list could be relatively low as the institutions usually target journals that are indexed or abstracted in specific databases. For instance, it would be impossible for an article to be published in a journal which is indexed in the Web of Science (WoS), if the institution targets the journals which are indexed in the Scopus. Consequently, the visibility of the published articles in the recommended journals could be relatively low.

Contacting the editors of potential journals is done by some of the authors to check the suitability of a submission. Usually, authors send the abstract of the manuscript to the editor to check the eligibility, while requesting more information that cannot be found elsewhere. This approach provides wider opportunity for the author to inquire about occasionally available information such as the journal's acceptance rate,

circulation, and review time. Despite the advantages of this approach, some practical difficulties could work against its reliability. For example, a short description about the study or the abstract of the manuscript will not provide enough details to the editor to decide the appropriateness of the manuscript for the scope of the journal. Confidentiality of certain types of information can also lead to avoidance of sharing information between an author and the editorial staff. Circulation statistics of a journal is not only unavailable for the public use, but also in most cases, the editorial staff has to request this type of information from the publisher of the journal. Even if, the editors and other editorial staff are enthusiastic to respond to the queries they receive, this method is not a consistent or reliable way to select a suitable publication outlet. Delayed replies from the editorial staff or not receiving a reply at all, can discourage the authors from considering the contacted journal in future submissions.

Reference lists of already published related articles can be considered as another possible approach for finding appropriate journals. The works cited by the author while preparing the manuscript also belong to this category. This simple and quick approach is more appropriate for identifying journals within the scope of the manuscript. Going further down, one can use the reference lists of relevant articles mentioned in the initial article to accumulate a list of potential journals. Therefore, this method can be used as a chain process to develop a sufficiently large list of candidate journals for submitting the author's manuscript. However, this method usually checks appropriateness solely by means of matching their subject scope, but not necessarily the author's publication requirements. Therefore, it may be necessary to refer a secondary source (e.g. website of the journal) to get a comprehensive idea about the target journal. Possibility of including discontinued journals or journals with recent changes of scope is another issue with a list of this nature. Nevertheless, this problem can be addressed to some extent by considering the date of publication of the corresponding article.

There exist several periodical directories which provide important metadata of schol-

arly publications. For example, Ulrich's periodical directory[5] and Cabell's directory[6] cover information of a considerable number of journals available at present. These publications are available online in addition to their print formats. Moreover, some prominent publishers (e.g. Elsevier[7], Emerald[8], Springer[9], and so on) facilitate researchers to retrieve their collection information according to a specified subject. As a result, an author can use these periodical directories or publishers' databases to find fitting journals by retrieving needed information from them. However, this approach requires access to subscribers only resources like Ulrich's and Cabell's directories. In general, metadata types and records included in these directories are sufficient to make a reasonable comparison among listed journals. However, since the directories include journals from a limited number of indexing databases, the subject diversity and the number of journals can be considerably low. Unlike subscription based directories, the information included in publishers' databases can be accessed freely. Availability of limited metadata types and unavailability of metadata for journals published by other publishers are the main practical difficulties of using publishers' databases. In addition to the above directories, a number of indexing services are available for finding information about journals. Journal Citation Reports (JCR)[10], SCImago Journal Rank (SJR) service[11], Scopus[12], and DOAJ[13] are some of the well-known indexing services which provide subscription based or free access to their contents. Limited number of metadata types and difficulty of finding information about journals apart from the indices they are listed in are the major issues with using them.

Getting assistance of commercial services is another possibility when deciding an appropriate journal. These fee levying services basically provide support for editorial

[5] http://ulrichsweb.serialssolutions.com/login
[6] https://www2.cabells.com/
[7] https://www.elsevier.com/
[8] https://www.emeraldinsight.com/
[9] https://www.springer.com/
[10] https://clarivate.com/products/journal-citation-reports/
[11] https://www.scimagojr.com/
[12] https://www.scopus.com/home.uri
[13] https://doaj.org/

works of scholarly works, but also extend their expertise to selecting suitable journals for an additional service fee. For example, Edanz[14], Enago[15], Editage[16], and editEon[17] provide this service for differing charges. Their suggestions may contain a list of potential journals sorted from the most suitable to the least. In general, authors are allowed to input their specifications of the desired journal and the service provides their suggestions accordingly. Their reasons for selecting and ranking corresponding journals can be expected from these commercial services in addition to the pros and cons of each journal. One significant feature of this approach is the quick processing time owing to the service fee. Despite the possibility of having a ranked list of journals with minimum effort, the main reason scholars don't flock to this could be the relative high cost of this method.

Methods included so far in section 1.4 discuss non-automatic approaches of journal selection that are generally more labor-intensive and time consuming.

Applications of information systems have initiated a few efforts to develop journal recommender systems to assist authors. Technical aspects of existing systems vary from simple factor based filtering, which decides appropriateness based on author specified subjects and submission requirements to complex content-based filtering, which considers textual aspects of a manuscript. Hybrid systems that combine both techniques have also been developed to minimize limitations and combine the advantages of using a multitude techniques. Minimum processing time, unbiased results, ability of comparing large number of journals and/or comparing across different journal databases, and OA facility are the major advantages of using these systems. However, these systems have varying capabilities based on technical aspects deployed in them. For example, the methods used to filter submission requirements and algorithms used for comparing text similarities could influence the accuracy of predictions significantly. The current study is targeted at this particular area of journal selection problem. More information about existing systems, their technical details, and

[14] https://www.edanzediting.com/
[15] https://www.enago.com
[16] https://www.editage.com/
[17] www.editeon.com/

advantages or disadvantages of using them can be found in sections from 2.2 to 2.6.

1.5 Recommender systems

With the rapid commercialization of the world, the use of recommender systems in industry has increased. Recommender tools are widely used for suggesting fitting movies, music, books, news, supermarket items, and scholarly works too. Ricci et al. (2015) define recommender systems as software tools and techniques which suggest most likely items of interest to a particular user. The predictions usually provide a measure of each item's importance with respect to a particular user's interests. Based on these measurements, system ranks a set of items from the most suitable to the least, which makes it easy to pick the best or settle for a lesser item based on a particular user's interests and available resources. Further, the process helps the user to minimize the confusion of selecting an appropriate item from a large set of similar items. Prominent websites, such as, Amazon.com, Netflix, Facebook, YouTube, Booking.com, IMDb, LinkedIn, ResearchGate, and Last.fm use recommender systems to extend and customize their services to users.

Recommender systems are beneficial both for the users and the service providers in numerous ways. Assisting in selecting the best matching item is the most important aspect of a recommender system, which is useful for users. In addition, a user can help other users in the community by contributing more information to a recommender system. For example, evaluation of items based on a rating system could help another user to make a more accurate decision. However, users must be cautious when making their decision as there could be malicious users who deliberately use such systems to promote certain items which would affect the accuracy of the recommendations. Service providers can get benefits from incorporating recommender systems into their interfaces as these recommendations can increase their sales. Possibility of identifying plenty of potential customers from all over the world and ability to understand their specific needs would help to diversify their range of items for the market. Accurate

and efficient recommender systems is useful to improve user satisfaction in addition to the services they provide. Finally, studying the patterns of user feedback is useful for the service providers to understand their needs more precisely. This factor is crucial for the stable progress of a service provider.

Burke (2007) identifies five different types of recommender systems, according to the factors they are using to determine user's profile and requirements.

Content-based: Recommendations are based on features associated with items and the ratings users allocate for them.

Collaborative: These systems match peer users' rating histories with current user's rating profile to make recommendations.

Knowledge-based: These systems match user's needs and preferences with existing features of the item.

Demographic: Demographic information about the user is utilized by these systems to make predictions. Suggestions may appear as a combination of ratings made by users in a particular geographic niche.

Hybrid: Different types of recommender systems have different strengths and weaknesses due to the different inputs they receive. For example, knowledge-based systems may work effectively even with relatively fewer information, while content-based or collaborative-based systems require more information to provide better results. Therefore, to enhance the performance of a system, one can use a combination of different types of recommender systems, termed "hybrid systems".

The current study proposes a hybrid system consisting of a content-based component and a knowledge-based recommender component. Thus, sections 1.5.1 and 3.4.1 provide an introduction to the concepts and tools, which will be utilized by the current implementation.

In addition to the authors, there are other users who can benefit from journal recommender systems. Some profit oriented corporations associated with editorial and

expert scientific review services use recommenders to find appropriate journals to publish their customers' articles. However, most of the times, this service is available for a fee which is not covered by the editing fee of the article. For instance, the Edanz journal recommendation service has an advanced paid recommendation procedure apart from their free, yet basic journal recommendation system. Although the Edanz advanced method is not fully automated, it gives a comprehensive analysis about the most appropriate journal outlet for a given manuscript. Hence, it can be seen that the possibility of deploying journal recommender systems is expanding from academic works to commercial purposes. Another party who could use recommender systems are the journal editors. Once the editorial board receives a manuscript, a recommender system can be used to get an approximate idea of the scope and discipline of the article. Inspecting the results suggested by the recommender system, the editors would be able to estimate the topical matching of the article to their journal. This practice can be used to filter manuscripts before starting an in-depth review. As a result, the massive amount of editorial workload can be reduced to some extent.

1.5.1 Concepts

Content-based filtering method has the advantage of user's rating profile along with item features to recommend appropriate items. This approach analyzes a set of descriptions of items rated by a user in a previous occasion and develops a model of user's interests using the features of items rated by the user. Then, items' features will be compared with the user's model and recommendations are made based on their compatibilities. These compatibilities are interpreted as relevance judgments between the user and items. The items with better relevance levels are recommended as the most suitable items for the user.

In general, content-based filtering needs three basic components to complete the recommendation process (Adomavicius and Tuzhilin, 2005). Content analyzer is the component that initiates the content filtering process. It pre-processes information

received from an input source. For example, extraction of feature information from the input source and making them ready for the next component of the filtering system are some major functions carried out by the content analyzer. The second component, profile learner, collects information of user preferences to build the user model. This model is based on the ratings the user assigned for the items in the past. The last component in the content-based filtering process, namely the filtering component compares user model with item's features to recommend appropriate items. As mentioned above, these judgments may take the form of relevance judgments and use algorithms to compute the similarities between the profiles of users and items.

The content-based recommender system component developed in this study includes some of the essential components mentioned above, though the profile learner does not work exactly in the same way as explained above. In the present system, the user profiles are not built by collecting rating histories of users. Here, the suggestions are customized more or less based on specific inputs (i.e. part of article texts), submitted by the users to the system. The dynamic nature of input contents from one to another does not work effectively when considering a more general scheme like user's past rating history. Therefore, instead of building rating profiles of users, the corresponding component of the current system only learns features of the input texts. Moreover, the present research utilized five algorithms to implement the filtering component of the content-based recommender system. These algorithms are responsible for identifying similarities of input text and training examples stored in the corpus to recommend the most comparable set of training examples to the input text. A brief account of each algorithm is provided in section 3.4.3.

1.6 Research questions

Productivity of a researcher rests upon a number of attributes. Attitudes towards social responsibility, intellectual capacity, capability of concentrating, and communication skills are only a few from a lengthy list of qualities that a productive researcher

should be endowed with. The responsibility of cultivating these traits greatly depends on the researchers themselves. This could be the common feature of these characteristics. In addition, availability of resources, motivation and proper guidance can be considered as another set of characteristics, which makes an external impact on the productivity of a researcher. This dissertation mainly discusses an electronic resource, which promotes the productivity of researchers as authors of scientific literature. Moreover, this novel approach is expected to have an external impact on a researchers' productivity. Bibliometric characteristics of journals in different subject domains could differ drastically from one another owing to the inspiration of different research cultures. Recognizing the most influential factors in a concerned subject domain is important to accomplish the researcher's professional ambitions within a quicker time span. Developing a recommender system as a resource to assist authors in publishing must concentrate on multiple issues. Composing a corpus with appropriate training documents and ensuring the effectiveness of the implemented algorithms are the foremost concerns of this process. Combining different methodologies reported in literature for recommender systems is another outlook taken by the present study to enrich the functioning of the system. Hybrid recommender systems approach is applied to perform functions which cannot be accomplished by a single type of recommender component alone. For example, the current research proposes to utilize a knowledge-based recommender component since the content-based recommender component cannot compare the bibliometric requirements of the target journals an author expects. The importance of individual recommender components can vary from one author to another based on the features they prioritize. To illustrate, authors who prioritize journals with similar content to the compiled manuscript may not be much concerned about bibliometric characteristics of a journal. However, this argument does not diminish the significance of coupling multiple recommenders together as there could be circumstances in which the converse of the preceding argument is convincing. As a final point, evaluating the proposed recommender system's performance against competitive gold standards and assessing to what extent the authors are benefited from the system is imperative

to have a rational idea of the usefulness of the current study.

The current research aims to address the aforementioned issues, while acknowledging the following major research questions:

1. How does the importance of the selection factors of OA journals differ in the two subject domains - medicine and social sciences?

   Journal selection factors and their importance can be changed from one subject domain to another, because of different research cultures followed by them. For example, medicine journals usually have higher journal impact factor values than the journals in social sciences. Moreover, differences between OA and non-OA journals could influence the author's journal selection decision. For instance, some OA journals charge an author fee to provide free access to published articles.

2. What is the most effective algorithm for a content-based recommender to determine the most appropriate journal for a given article abstract in each subject domain?

   Behavior of distinct similarity algorithms could vary based on the characteristics of the text documents they compare. For example, length of documents and frequency of similar terms could influence the performance of the algorithm. Also, the document characteristics in different subject domains are likely to depend on the considered domain.

3. Does the knowledge-based recommender component significantly improve the performance of the content-based recommender component?

   Hybrid journal recommender system combining a knowledge-based recommender component for OA journals with a content-based recommender component could increase the performance than using the content-based recommender component alone. However, it is important to understand to what extent the performance of the hybrid recommender system can be improved than the content-based recommender component, if the hybrid recommender

outperforms the content-based component. This would reveal whether the knowledge-based recommender component adds valuable information to improve the hybrid recommender system.

4. Does the new journal recommender system offer appropriate suggestions with respect to gold standards and authors?
   This research question evaluates the final outcome of the study. It compares the performance of the hybrid recommender system with benchmarks. Moreover, understanding the performance of the hybrid journal recommender system from the author's point of view is essential since they are the major target group of the current research.

To address the above research questions, the current study proposes a new journal recommender system. The new recommender system is developed based on four major stages as follows:

1. Identifying and prioritizing author's journal selection criteria in general.
2. Developing a content-based recommender system with an appropriate algorithm.
3. Collecting journal metadata and developing a knowledge-based recommender system.
4. Configuring and evaluating the performance of the hybrid journal recommender system.

1.7 Organization of the dissertation

The first chapter of the dissertation establishes the background of the current study. This chapter describes the progression of publishing and the significance of academic publishing for researchers. Comprehending the issues faced by authors throughout

the publication process is important to appreciate the aims of the current study. Significance of medicine, social sciences, and open access journals is described to justify their use in this dissertation. Established as well as more recent methods that exist to minimize the challenges of publishing are also examined along with their inadequacies. Further, a gentle introduction to recommender systems is given as this approach can be considered as a relatively novel method that aims to assist authors in finding appropriate publication outlets. Finally, chapter 1 highlights the major research questions of the study.

The second chapter of the dissertation discusses studies related to the current research. The chapter commences describing numerous journal selection schemes developed in previous studies. Most of these studies have investigated publication factors while giving special attention to a particular subject domain, except for a few generalized schemes. In addition to publication factors, both graphical and mathematical models developed using these factors are examined in the second chapter. Thereafter, some of the established content-based recommender systems including JANE, eTBLAST, IEEE publication recommender, and Elsevier journal finder, and their distinctive features are reviewed. Afterwards, hybrid journal and conference recommender systems that merge content-based systems with collaborative-based systems are outlined. Apart from well-known filtering methods like content or collaborative, lesser known methods such as log data analysis and methods using simple bibliometric data are introduced in the chapter as alternative approaches for journal recommendation. To conclude the chapter, differences between the proposed method and the existing journal recommendation methods have been elaborated on, in order to justify the significance of the current study.

The third chapter of the dissertation is devoted to illustrating the methodology followed in the current research. The chapter commences with the first literature survey conducted to identify the important factors, to determine the most appropriate journal outlet for publishing. The manuscript submission considerations survey of authors (hereinafter this survey will be referred to as the first author survey)

targeted at finding the weights of importance the authors assign for each factor is described in detail thereafter. Next, the content-based recommender system developed by the current research is elaborated giving a specific emphasis on selecting the most appropriate similarity algorithm for the two separate subject corpora. In addition, similarity algorithms tested by the current study and the software tools used to implement the algorithms are described, enriching the background information further. The chapter introduces the knowledge-based recommender component. Collection of journal metadata and the measures employed to assess author's journal selection criteria and the journal's available criteria are discussed in addition to its software implementation. The survey for collecting data to configure the recommender system (hereinafter this survey will be referred to as the second author survey) is then described. This survey collected journal selection criteria considered by a sample of authors when submitting one of their recently published articles. The accumulated information from this survey will be used in configuring the new journal recommender system. The mapping criteria utilized to convert the categorical answers received for survey questions into numerical form is then illustrated using a few examples. To conclude the chapter, an author survey for evaluating the journal recommender system (hereinafter this survey will be referred to as the third author survey) is presented. This third author survey was designed to return the lists of suggested journals based on the answers received for the second survey.

The fourth chapter is devoted to representing the results achieved by the current study. This starts off with the results obtained for the first author survey. One of the most important findings as described in the chapter is quantifying the weights of importance of the 16 journal selection factors for the two subject domains independently. In addition, the factors that attain significantly different importance in the two subject domains and important correlations between factors are communicated as other crucial findings of this survey. The results are also published in Wijewickrema and Petras (2017). Next, the chapter represents the performance results of the five algorithms employed to implement the content-based recommender system. Performance of the algorithms is compared with respect to the individual

test cases as well as the sub-disciplines. Results accumulated for the second author survey are also illustrated in the chapter. In essence, the results disclose how far the author's publication expectations in terms of the bibliometric factors of journals are fulfilled by the journals they are actually publishing in. At its conclusion, the chapter announces the third author survey, requiring the respondents of the second author survey to rank the journals suggested by the system from their perspective. Survey results are analyzed and the performance of the final recommender system is assessed in addition to comparing its performance with the exclusive use of the content-based recommender component.

The final chapter of the dissertation provides the conclusions obtained in the current study. The chapter represents conclusions along four major themes. To begin with, the conclusions of the first author survey are deliberated. This includes potential reasons for exhibiting significantly different importance by some publication factors in two subject domains, besides imparting the importance received by each factor in the two domains individually. Second, the conclusions relevant for opting for the most appropriate algorithm for implementing the content-based recommender system are elaborated. This part moreover examines the performance disparities of the algorithms employed assessing their distinctive qualities. Conclusions with regard to the results of the second author survey are depicted consequently. Conclusions regarding the disposition of the sample authors are illuminated in addition to conferring the likeness amongst the expected journals and the actual journals, in which the authors published their articles. Subsequently, the conclusions of the recommender system's assessment are communicated in the final chapter. Afterwards, the performance of the final system is evaluated with the gold standards and against the content-based recommender component. Thenceforth, the significance of the current study is substantiated. This section ultimately aspires to scrutinize the innumerable users who can profit via the revelations of the research. Further, it particularizes the means of the results of the first and second author surveys, and the final recommender system can be exploited by the users. Last section of the chapter is dedicated to the viable future directions of the current research. The majority of the indications in

this section focus on enhancing the current recommender system exploiting technical aspects. Besides, potential augmentations to first and second author surveys are advocated to discover the rationale behind certain authors' behavior.

Chapter 2

Literature Review

"Every great advance in science has issued from a new audacity of imagination"

– John Dewey: *The Quest for Certainty* (1929)

Section 1.1 of the dissertation presents a brief introduction to academic publishing including its objectives, characteristics, and usefulness. This chapter imparts the previous efforts in finding fitting publication outlets for manuscript submission by exploring along four major areas, namely: establishing journal selection criteria and models, content-based journal recommender systems, hybrid systems including a collaborative-based recommender component, and unconventional journal recommendation methods. Figure 2.1 summarizes various publication outlet selection approaches that are described by the current chapter.

2.1 Criteria and models developed for journal selection

The studies which are discussed under this topic did not attempt to implement a precise journal recommender system for publication outlet selection. They do however, essentially recommend a multitude of criteria based on multiple factors to utilize as the benchmark to shortlist the available possibilities and to select the best publication outlet for publishing. Furthermore, publication factor identification laid the foundation for most of the selection models and automatic recommender systems discussed in the rest of the chapter. This paragraph outlines prevailing literature

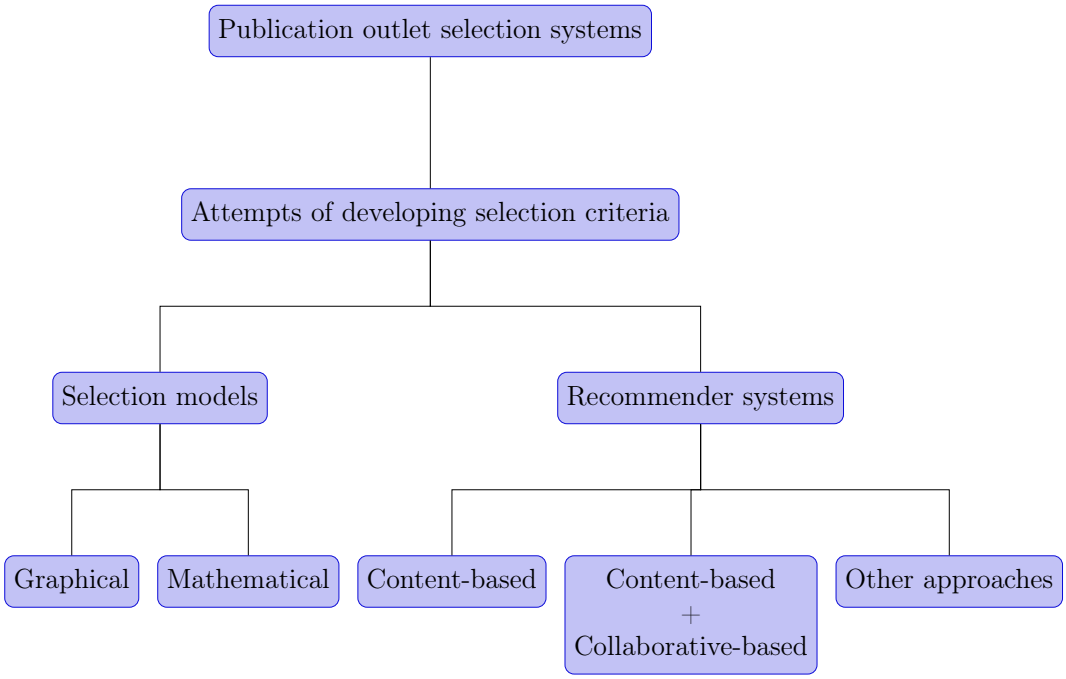


Figure 2.1: Different publication outlet selection methodologies

on journal selection limited to a particular discipline. However, the literature in the next paragraph discusses the journal selection process without restricting to a particular subject domain. Bröchner and Björk (2008) studied author choices in construction management journals involving 397 authors with 35% response rate. This

work concentrated particularly on authors' perception regarding the quality and the service of the journals. It has collected data based on two schemes. The general scheme concentrated on nine aspects independent of the author (e.g. likelihood of acceptance, relevant readership) while the second scheme was based on author's experience (e.g. career value, helpfulness of review). The survey results revealed high academic status of the journals as the determining element for authors when deciding the most appropriate journal. Moreover, the respondents considered its free availability with the least amount of interest. Cheung (2008) surveyed a sample of authors selected from five prestigious educational journals. The study used 24 factors to identify their level of importance for the researchers in the education field. Among the other factors, outstanding contributors to the journal, knowledge of editor's intellectual interests, and previous rejections by the journal were some uncommon factors investigated by the study. Moreover, it was reported that the authors of the education field place great emphasis on readership and topical resemblance than other factors. A global email survey organized by Søreide and Winter (2010) concerning 15 arbitrarily chosen factors relevant to selection strategy of surgical journals found the reputation of the journal as the most influential factor for the majority in the field, followed by the IF. The least important concerns included the acceptance rate, possibility to suggest peer-reviewers, and OA facility. Although the study reckoned the future safeguard of the traditional paper-based surgical journals due to the less attention of authors towards OA publishing, it contradicted the former conclusion of Rowlands and Nicholas (2006) who affirmed the potential disruption OA could bring about for traditional publishing. Özçakar et al. (2012) reported the results of a survey conducted to identify the publication factors of physiatry. The study scrutinized 20 journal selection factors and found that bibliometric indexes (except IF) have the lowest rank. According to the authors, the novelty of bibliometric indicators and their uncertain validity may lead to this negligence. Further, they discovered that the area of interest of a journal as the highest ranked aspect. Rousseau and Rousseau (2012) investigated authors' perceptions about journal selection from the information science view point. They suggested 12 factors to rate; the general stand-

ing of the journal entertained the highest attention and its importance surpassed the importance of IF as well. Surprisingly, unlike in most of the other studies, topical match earned the least importance according to the results. A journal selection criteria developed by Shokraneh et al. (2012) employing focus group conversations, feedback of workshop participants, and non-systematic review of relevant literature uncovered 14 major factors of interest with regards to biomedical authors. The paper informed communication with the journal as a new factor to consider besides more frequent determinants like manuscript topic, prestige, quality, cost, and so on. The significance of a user-friendly website, authors' freedom to suggest reviewers, email alerts, and quick responsiveness are described as the qualities of a good communication practice. Bröchner and Björk (2008) have investigated all the factors covered by Björk and Holmström (2006) excluding the scientific level of the journal, but included a few other factors which were not examined in the previous research. Their study though was restricted to the library and information science discipline. By means of a web-based questionnaire, 326 participants were invited to rank 21 journal selection factors. The two factors, peer-review and topical fit attained the highest author attention according to the outcomes of the research.

Apart from the studies focusing on specific disciplines, some have reported journal selection studies over a diverse range of fields. A large-scale survey by Rowlands et al. (2004) using an initial sample of 107,500 authors has proven that the right readership is the most influential factor whereas coverage by abstracting and indexing services, IF, and the composition of the editorial board follow close behind. The conclusion emphasized that the importance of the last three factors do not differ widely. The price of the journal was the least important factor for the selected sample. Finally, this study reckoned the authors' resistance to update their publishing aspects from current to a new status and therefore, the minimum likelihood of altering their conservative notions about OA publishing within a short term. Pursuing the same methodology of this research, but covering a broader area of publishing was conducted by Rowlands and Nicholas (2006). The research examined the authors' attitudes and perceptions regarding a range of issues they face by a scholarly

communication system. As a part of their questionnaire, they incorporated 10 journal selection factors to decide how seriously authors deem them. Reputation of the journal, readership, and IF in that order, ranked as the top three most important factors, while copyright was rated as the least worthy factor. The value attached to peer review status was examined by the survey separately and it was learnt that almost all respondents indicate it as a very important or quite important factor to be considered. Remarkably, the respondents believed that downloads can be established as a more effective measure of article usefulness than citation counts. As a final point, the respondents did not entertain strong positive attitudes towards OA publishing and they thought a major shift to OA publishing could interrupt the current publishing system, confirming again the earlier conclusions regarding OA publishing by the same authors (Rowlands et al., 2004). Solomon and Björk (2012) publicized the results of a survey particularly engaging a sample of OA authors. The sample represented 1038 authors from seven independent disciplines and explored sources of funding for OA publishing and six potential influencing factors in choosing journal outlets. Proper fit of the manuscript with the journal scope, quality, and publication speed of the journal scored the highest marks for importance while readership type, OA status, likelihood of acceptance were rated as unimportant. One can also find a few more accounts on journal selection, which were not based on formal author or literature surveys, but on the experience of working in the field. Broome (2007) addressed the importance of two factors, namely, degree of editorial control and availability of special issues or supplements which received no attention of the authors until then. Lewallen and Crane (2010) advised considering multiple factors organized under four major themes: appropriate audience, topical fit, purpose, and coping with the journal guidelines before submitting a manuscript. The significance of the proposed criteria is that it compares the qualitative aspects between the manuscript and the targeted journal than considering quantitative metrics. Sharman (2015) discussed 23 aspects that could influence the author's choice of the journal. In addition to the factors which have previously been discussed in literature, some uncommon factors such as the submission deadline, publication policy of the research funder,

and publicity were recommended as important concerns. According to the studies mentioned in the current and previous paragraphs, there exists no consistency in the significance of the journal selection factors from one study to another. This cannot be explained by the variation of the examined factors in different studies since most of the studies had some shared factors. Therefore, the reliance of publishing needs on the specific subject area the authors work in could be a possible reason for this behavior.

In addition to identifying the factors that influence the author's decision of a suitable journal outlet, some studies extended their attempts to proposing journal selection models too. These models can be divided into two parts: graphical models and mathematical models.

Björk and Holmström (2006) reported a study with eight factors which directly influence the author's decision of journal selection. In addition to major factors, the study also describes 21 other underlying factors. They have used all these factors to develop a selection model, consisting of four blocks namely, infrastructure, readership, prestige, and performance. Each block of the model included a number of decision making factors according to their relevance to the block. The authors also suggested a number of methods to collect data for measuring different factors of the model. For instance, gathering directly available data, calculating data based on available information, and gathering data from publishers are some of the methods they have suggested. This model and proposed journal selection factors were reused in a later study (Björk and Öörni, 2009) to test the method for three sets of journals from different fields. The findings disclosed existing problems with data acquisition methods proposed by the previous study. Knight and Steinbach (2008) identified 39 different journal selection considerations spread over three primary dimensions namely, likelihood of timely acceptance, potential impact of the manuscript, and philosophical and ethical issues. Although the study initially intended to examine the domains of information systems and information science, the ultimate goal was to cover all authors regardless of the discipline. This paper discussed 11 aspects

in detail to make a fair decision about journal's prestige, as most of the previous studies have given substantial attention to this factor. Also, the study split-up the factor – publication time, into two separate components: review cycle time delay and publication time delay. A graphical model was proposed as a two-dimensional grid with the axes, likelihood of timely acceptance and the potential impact of the manuscript. The third dimension, philosophical and ethical considerations works as an outer wrapper for the previous two-axis grid. Wijewickrema (2015) proposed a three dimensional model for selecting the most appropriate journal based on 40 publication factors. The study recognized potential factors based on a literature survey. According to the findings, 10 publication factors were identified as important by more than 50% of the previous studies concerned. The dimensions of the model are identified depending on the way the factors increase or decrease the relevancy between the journal and the manuscript.

The current paragraph illustrates some mathematical models developed to rank journals for selecting appropriate ones based on citation-based indices and some other selection factors such as fast review time, acceptance rate, speed of publication, and so on. A careful observation reveals that they have not considered more than two or three publishing factors to construct the models. The difficulty of controlling the model with an increase in the number of factors could be the closest reason for this. One of the earliest mathematical models was proposed by Oster (1980). This numerical approach represents an optimal order for submitting manuscripts from a list of pre-determined journals based on potential benefits. Quick reviews, relatively higher acceptance rates and a few other factors have been considered as the benefits in this study. An extended version of this approach was developed by Heintzelman and Nocetti (2009) while reducing previous calculations considerably. He and Pao (1986) proposed an algorithm to rank a set of journals in a specific discipline. The suggested method compares the citations received by a set of recommended journals in the considered field from each candidate of the target journals' list. Then the list of target journals is ordered according to the scores generated by the algorithm based on the citation counts. Ease of implementation and the possibility to accom-

modate any discipline can be seen as the important features of this approach. A relatively simple, but dynamical journal recommendation algorithm was introduced by Gutknecht (2014). The proposed method generated an ordered list of journals according to the author's publication needs such as peer-review status, speed of publication, IF, rejection rate, and so on. The mean importance given by the authors for each publication factor was determined by a survey and they were incorporated to the final algorithm. In addition to the authors, funders, librarians, and institutions were also surveyed separately to know their interests regarding the publication factors, because the importance could depend on the service they offer. The algorithm returned the sum of the normalized numerical values earned by each considered publication factor. However, each factor value was multiplied by the corresponding mean importance as the significance of the factors differs with regard to authors, funders, librarians, and institutions. It was reported that the proposed algorithm was more effective compared with JANE and Edanz journal selector[1]. A composite journal indicator combining five standard indices: IF, SJR, h-Index, Source-Normalized Impact per Paper (SNIP), and Immediacy Index was derived by Bradshaw and Brook (2016). Ranked results of the new indicator were compared with the order arranged by experts in the field for sets of journals from ecology and multidisciplinary fields. Existence of a $0.68 - 0.84$ correlation between the datasets was reported. Its ability to compare journals within or among several disciplines was emphasized as one of the major advantages of this indicator. A more recent study by González-Betancor and Dorta-González (2017) proposed an alternative approach to both IF and h-Index, as they heavily depend on the discipline of the journal and the journal size. This new model considered the percentage of journal's highly cited publications as an indicator of the scientific impact. Five citation percentile categories were defined as 10, 20, 25, 30, and 40 and compared with the citation distributions of IF and h-Index for two separate time windows over four separate disciplines. Results showed that the new indicator shows relatively higher homogeneous citation distributions among different disciplines and, therefore, is more suitable to compare journals across disciplines. As

[1] https://www.edanzediting.com/journal-selector

a closing note, this paragraph does not include regularly used journal ranking models like IF (Garfield, 1972), h-Index (Hirsch, 2005), Eigenfactor score (Bergstrom, 2007), and SJR (González-Pereira et al., 2010) as they have often and intensely been discussed in literature elsewhere. All these indices are primarily based on citation counts and journals are ranked by just one factor.

Table 2.1 summarizes the detailed information included in section 2.1. The complete list of factors collected by the current study, including the factors in table 2.1 are given by figure 3.2.

Article	Subject domain	Collecting factors	No. of factors
Bröchner & Björk (2008)	Management	Author survey	11
Cheung (2008)	Education	Author survey	24
Søreide & Winter (2010)	Surgery	Author survey	15
Özçakar et al. (2012)	Physiatry	Author survey	20
Rousseau & Rousseau (2012)	Information science	Author survey Focus group	12
Shokraneh et al. (2012)	Biomedicine	Workshop feedbacks Literature survey	14
Bröchner & Björk (2008)	Library science	Author survey	21
Rowlands et al. (2004)	Multidisciplinary	Author survey	10
Rowlands & Nicholas (2006)	Multidisciplinary	Author survey	10
Solomon & Björk (2012)	Multidisciplinary	Author survey	06
Broome (2007)	Multidisciplinary	No formal method	02
Lewallen & Crane (2010)	Multidisciplinary	No formal method	04
Sharman (2015)	Multidisciplinary	No formal method	23
Björk & Holmström (2006)	Multidisciplinary	No formal method	08
Knight & Steinbach (2008)	Multidisciplinary	Literature survey	39
Wijewickrema (2015)	Multidisciplinary	Literature survey	40
Oster (1980)	Economics	No formal method	04
He & Pao (1986)	Multidisciplinary	No formal method	01
Gutknecht (2014)	Multidisciplinary	Stakeholder survey	19
Bradshaw and Brook (2016)	Multidisciplinary	No formal method	05
González-Betancor & Dorta-González (2017)	Multidisciplinary	No formal method	01

Table 2.1: Literature summary of identifying factors, building graphical, and mathematical models

2.2 Content-based journal recommender systems

eTBLAST (Wren et al., 2007) is a content-based journal recommender tool which can find the appropriate journals, the authors with expertise knowledge in a given

field, and articles fitting with a given query. This system supports retrieving suitable records basically from the MEDLINE database. In addition, it also searches databases like NASA, arXIV.org, RePORTER, and other similar ones. Instead of searching subject terms, eTBLAST is able to decide the similarities between the database records and a given article by using its abstract. It extracts and analyzes weights of keywords contained in the submitted text to identify the related records in the considered database. Then a sentence alignment is performed to obtain a similarity score to determine the relevance. However, an experiment using 4230 abstracts proved that only in 33% of cases, eTBLAST ranked the journal in which the abstract was published within the top 10 suggestions (Wren et al., 2007).

JANE is a freely available web-based application which mines MEDLINE medical database to find appropriate journals for submitting manuscripts or to find potential reviewers among competitive peers (Schuemie and Kors, 2008). Moreover, it helps researchers to retrieve similar articles to a given input. JANE allows authors and editors to enter the title and abstract of the article for which they need to find an appropriate journal outlet, a reviewer, or a similar research paper. As a result, the system returns a list of potential journals, authors, or articles according to a rank order of their suitability. This tool follows a special technique to build the confidence of the authors who submit their abstracts. Since some authors may be reluctant to submit their novel findings to an unknown system, JANE supports scrambling the input text and arranging the words in alphabetical order. This process makes it difficult to capture the original input text. The implementation of JANE is based on Lucene search engine library which is an open-source tool. It uses Lucene's MoreLikeThis algorithm (Johnson, 2008), which is based on the well-known Vector Space Model (VSM) to determine the similarity between the input text and the documents in the database. This produces an ordered list of 50 articles according to their relevance with the input text. Then the system uses a weighted k-nearest neighbor approach to generate a list of journals based on the retrieved articles. The similarity scores of all the articles belonging to each journal in the list are summed and normalized separately to determine the 'confidence' of each journal in the retrieved list. Then

the final ordered list based on these confidence scores allows JANE to suggest the most appropriate journal for the input abstract. The same procedure is used by JANE to determine appropriate reviewers too. Although, VSM includes a variety of similarity measures, it is difficult to find the implementations of most of them on journal recommender systems. This can be seen as a gap in the present recommender systems research since there could be slight performance variations among the similarity measures which belong to the same Information Retrieval (IR) model.

eTBLAST and JANE are not the only recommender systems which can be used to suggest appropriate journals. There are several commercial and non-commercial solutions with different facilities. Some of these systems are based on MEDLINE as JANE does. Furthermore, one can find publishers who allow using their recommender systems to search suitable journals from their own databases. Another category of the journal recommender systems facilitates users to discover fitting journals via cross database search. However, this is also limited to only a few sources.

The Edanz Journal Selector is a web-based journal recommender tool, which uses advanced search algorithms and natural language processing techniques according to their official web page (Gutknecht, 2014), yet further information on their algorithms is not published. Their journal corpus is built by collecting abstracts and articles from multiple sources including PubMed and Springer. The tool supports scholars in refining the retrieved list of journals based on the IF and the publication model (i.e. open access, hybrid, or non-open access). Moreover, Edanz services offer a fee-based journal selection service in addition to their free, online journal selection system. The fee-based service uses experts' publication knowledge to suggest a suitable journal for submitting a manuscript (Rison et al., 2017).

The database of online tool called JournalGuide includes more than 40000 article metadata records from PubMed, CrossRef and from several publishers and aggregators (Mudrak, 2015). This journal suggester allows to input an abstract and title separately to suggest suitable publication outlets. The ordered list is based on a score, but also allows comparing the suggestions based on the publisher, IF, publi-

cation speed, and OA status. However, one has to consider these factors separately from each other, as a composite score for them is not calculated. The system uses a proprietary algorithm to make comparisons between input text and database documents though more details of the mechanism have not been revealed. Similar to Edanz, this service allows authors to submit their works for a peer review process. Then a list of customized journal recommendations is given based on independent evaluations of three academic experts in the field.

As a supporting tool of a larger and well established academic network, the ResearchGate Journal Finder helps researchers to find the best fitting journal to publish their manuscripts (Tattersall, 2016). This journal locater allows users to copy and paste the abstract of a manuscript to find an appropriate journal outlet. However, the mechanism behind text matching is not documented (Gutknecht, 2014).

Rollins et al. (2017) introduced a journal recommender system, which is known as the Manuscript Matcher. This tool uses both Support Vector Machines (SVM) and k-Nearest Neighbor (kNN) algorithms concurrently to compare the suitability of the recommended journals to the author's manuscript. The average confidence score of both algorithms is considered to select the most appropriate journal outlet. The system is further enhanced by filtering results based on bibliometric elements. Full-text articles for the corpus are collected from various OA repositories while their meta-data records are taken from the WoS. Therefore, the corpus documents are ultimately limited to journals indexed in the WoS. In order to evaluate the system, the authors have completed a user survey and received 64% positive feedback from the users. Manuscript Matcher is developed as a commercial product associated with the EndNote[2] reference management system. Lack of specific range of bibliometric elements for filtering selected results can be seen as a major drawback of the existing system.

Some of the existing journal recommender tools restrain their suggestions only to their own journal databases. Thus, the author's ability to see a wide range of po-

[2]http://endnote.com/

tential publication options is reduced when using them. This is a common practice of most the commercial publishers who provide recommender tools associated with their journal services. Further, this strategy enables publishers to promote their own journal collections. As a publisher, Elsevier supports their users in finding suitable journals as well. Their tool[3] matches around 2900 journals from Elsevier's ScienceDirect database. The system permits inputs of the title and abstract of the article separately to perform the text based selection. In addition to textual similarity, it allows to refine results according to the subject of the article and uses a field-of-research specific vocabulary. Furthermore, the system permits to restrict the results to OA journals. The Okapi BM25 algorithm has been implemented as the similarity measure of the system (Kang et al., 2015). Therefore, the results are minimally biased from the lengthy documents in the training database. This system first analyzes the field-of-research specific vocabulary with the input text by author and converts the result into a numerical figure called fingerprint. Then it is compared with the fingerprints of all the articles in the database and 10 matching journals are suggested based on the fingerprint similarities. IEEE Publication Recommender[4] allows matching articles in the field of technology (Forrester et al., 2017). The system covers nearly 200 periodicals and 1500 conferences, while facilitating searching of both or one source at a time separately. Unlike in most other systems, IEEE tool accepts even full-text article uploads to find appropriate journals beside article abstracts, titles, and keywords. In addition, one can refine search results according to the date the article is expected to be published. However, the system facilitates the comparison of articles from only IEEE publications and ranks the results by relevance. Springer Journal Suggester[5] is a free web-based tool, which allows authors to find a fitting journal from more than 2500 of the Springer and BioMed Central journals (Alhoori and Furuta, 2017; Rollins et al., 2017). As usual, separate entries of title and abstract are used to retrieve a list of appropriate journals. Nevertheless, its refinement criteria deviate from other systems as it allows to specify indexing

[3] http://journalfinder.elsevier.com/
[4] http://publication-recommender.ieee.org/home
[5] https://journalsuggester.springer.com/

services in which the expected journal must be included. Further, minimum margins of IF, acceptance rate, and time to first decision are the other refinement options that authors are provided with. However, the semantic technology – the technology the system uses to make text comparisons is not available as in most commercial publishers. Wiley journal finder is a new author service initiated by the publisher to assist authors in selecting fitting journals. The application improves the input text using Luxid[6] – a semantic content enrichment platform and then performs a similarity calculation based on the data retrieved from Wiley Knowledge Store, where numerous data of articles published by Wiley is stored. The search results are displayed with categories assigned to journals according to the Wiley subject taxonomy, offering authors a comprehensive understanding about the specific scope of the suggested journal. One of the major drawbacks of the system though is its inability to find meaningful results for queries with less than 1000 characters. Hence, article titles are not always sufficient to obtain good results from the Wiley journal finder.

2.3 Hybrid venue recommender systems with a collaborative-based component

The publication venue recommender system proposed by Pham et al. (2011) is an extension of the conventional collaborative filtering method. This paper emphasized the problems inherent to collaborative recommender systems due to data sparcity. To minimize this problem, the new system used a clustering approach based on the social information of the authors. The authors with similar research interests are clustered based on co-authorships. Then a conventional collaborative filtering algorithm was applied using clusters as neighborhoods. Precision-recall curves obtained after implementing the new method proved that the clustering approach performs better than the conventional cosine based collaborative filtering technique. Clustering authors based on reference information and participation in similar conferences

[6] http://www.temis.com

was suggested to enhance the social network further for better performance. Flexibility of adopting the theory for recommending journals in addition to conferences was one of the major advantages of this attempt.

Yang and Davison (2012) developed a recommender system targeting academic conferences, but not specifically for journals. This content and collaborative-based approach considered both topic and writing style information. The first part of the system identifies suitable publication venues by comparing similar articles published in previous conferences. This comparison is made by the cosine similarity measure. In addition, more than 300 distinct features have been used under three major aspects: lexical, syntactic, and structural. Finally, the system considers other papers authored by the author of the target paper, papers cited by the target paper, and papers that share similar citations with the target paper to decide the most appropriate publication venue. This novel system was tested against four baseline algorithms and its effectiveness beyond others was established. Further, the papers that are authored by the same author are identified as the most reliable criteria that improves the effectiveness of the system.

A conference recommendation method based on the information of author's publication network showed more effective results than a content-based recommender system approach using kNN algorithm (Luong et al., 2012). Their approach constructed a social network for each author in the corpus documents extracted from 16 conferences. The social network included information of each author's publications and co-authors. The recommendation process was based on the reputation of the author's social network. For example, it analyzed the most frequent conferences the authors in the social network published and the strength of the association between co-authors and the main author.

The system proposed by Boukhris and Ayachi (2014) was a hybrid recommender by nature as it combined a collaborative recommender engine with a community recommender engine and also a utility-based recommender component. The conferences authored by the researchers those who have already cited the works of the

target author were numerically valued under collaborative component while conferences represented by co-authors and the authors belonging to the same institution as the target author were assigned a score under community recommender component. The utility-based recommender was used to filter results further to match the target author's requirements like conference location, rank, and publisher. The system evaluation verified that the combination of community and utility-based recommender components improves the results over using the collaborative system alone.

2.4 Other journal recommender systems

Lu et al. (2009) reported a web application to search fitting journals through log data analysis of PubMed. This system facilitates authors to find suitable journals using subject terms and allows sorting the retrieved records according to their popularity. This popularity is determined by analyzing a journal's previous usage. For the evaluation, they have used 29 participants and asked them to compare the results given by their system arranged according to the popularity and the alphabetical order of the journals. As a result, all the users have recommended the list ordered by popularity. Moreover, a relatively higher precision (0.910) and recall (0.893) values were reported. However, this system tends to retrieve some highly diverse journals like *Nature* repeatedly because of their popularity.

Cofactor Journal Selector[7], does not perform a textual similarity matching to determine the suitability (McKiernan et al., 2016). Primarily, the matching is based on five specified needs of an author. These include the subject of the article, peer review policy, OA status, speed of the publication process, and other aspects like article length, IF, and copy-editing. The database of the system contains a limited number of journals mainly from biology and medicine. They are also limited to general subject scopes similar to those in mega journals. The inadequate journal options it provides and the restrictions in the selection criteria reduce its use even if

[7] http://cofactorscience.com/journal-selector

the recommendation process is automated.

2.5 Summary of literature

In the section 2.1 of the current chapter journal selection factors limited to a number of subject domains and based on multiple disciplines were identified. It was revealed that no specific factors were limited to certain subject domains, but the factors were common to all subject domains, more or less. However, the level of importance of factors considered in different subject domains varied from one domain to another. Author survey and literature surveys were conducted mostly for identifying journal selection factors and their importance. In addition to identifying factors and their importance, some of the studies developed graphical and mathematical models for selecting an appropriate journal outlet. The number of factors used for the development of mathematical models was very low compared with the other studies. Topical matching, readership, IF, journal's reputation, and publication speed were some of the factors which attained the highest importance, while factors such as OA status, acceptance rate, and copyright issues were gained the least attention of authors. A few uncommon journals selection factors such as journal's scientific level, publication policy, publicity, submission deadline, and so on was also introduced in this section. Overall, the journal selection factors introduced in section 2.1 could be considered as the foundation for most journal recommender systems described under sections 2.2, 2.3, and 2.4.

Section 2.2 of the chapter describes the existing content-based journal recommender systems. The section includes freely available systems such as eTBLAST, JANE, Ednz Journal Selector, JournalGuide, ResearchGate Journal Finder, Manuscript Matcher, Elsevier Journal Finder, IEEE Publication Recommender, Springer Journal Suggester, and Wiley Journal Finder. The two systems – eTBLAST and JANE mainly retrieve appropriate journal records from the MEDLINE database and limited their suggestions to the medicine subject domain. Edanz Journal Selector and

JournalGuide are based on multiple databases including PubMed. Other recommender systems described in this section are based on the databases maintained by commercial publishers such as Elsevier, IEEE, Springer, and Wiley. These publishers have their own journal recommender systems, yet limit their recommendations to the journals they publish. The similarity algorithms implemented in these systems were hardly found in the literature. Lucene's MoreLikeThis with kNN, SVM with kNN, and BM25 were utilized by JANE, Manuscript Matcher, and Elsevier Journal Finder respectively. In addition to textual similarities, some of the above journal recommender systems allow authors to filter results based on journal selection factors.

Section 2.3 discusses collaborative-based journal and conference recommender systems. However, some of the studies included in the section combine collaborative component with other recommender components such as content-based and utility-based. Improving the collaborative-based journal recommender approach using clustering methods and using different aspects such as author's reputation based on social network information and the citation based methods for collaborative approach were discussed in the section. Some of the recommender systems mentioned in this section are limited to recommending appropriate conferences, but methodologies could also be adapted for developing journal recommender systems.

Section 2.4 of the chapter illustrates other journal finding approaches including manual methods and recommender system approaches that were not discussed in the previous sections of the current chapter. For example, usage statics of journals from log data of databases has been used as a manual approach to find target journals. In addition, matching small number of factors with corresponding factors of the existing journals was introduced as a recommender system approach. These approaches can be considered as very simple approaches available for journal selection compared to other systems discussed in sections 2.2 and 2.3. However, these approaches likely to perform poorly than other systems due to the lack of selection factors and simplicity of the methodologies applied for comparing manuscripts with corpus documents.

2.6 Difference: proposed system and available systems

The current study contends that the proposed journal recommender system deviates from the previous studies which developed criteria for selecting appropriate journals. The new study established journal selection criteria for two subject domains and found relative importance assigned by the authors for each factor of the criteria. A recommender system using the established criteria and with some mathematical implementations was developed in the study, endeavoring to extend the criteria and models developed for journal selection described in section 2.1. Additionally, there is no other study mentioned in the literature that has completed a comparative study between the social sciences and medicine fields to evaluate the factors that influence publication outlet selection.

Recommending journals from large OA databases such as DOAJ is important for authors who expect to increase the number of readers of the published articles. Some of the journal recommender systems discussed above assist authors to select OA journals via cross-database searching, but not solely from OA databases. However, this is not sufficient due to two reasons. First, these recommender systems do not search larger OA databases and may fail to notice more appropriate OA journals for a given manuscript. Second, searching for OA and non-OA journals simultaneously would not be effective as some of the characteristics of OA journals widely differ from the non-OA journals. For instance, free availability of OA journals likely to increase the usage of OA journals than the non-OA journals. Therefore, applying the same values for these factors could not generate precise recommendations for both of these two types of journals. The current journal recommender can avoid these limitations since it targets to recommend only OA journals.

Most of the content-based journal recommender systems discussed in section 2.2 used similarity measures different to the current study. Moreover, it was not reported whether the performance of more than one similarity measure for their corpus was

tested in those studies. This can be considered as a major weakness of their study design as the performance may well depend on the nature of the corpus documents. In addition, except the IEEE Publication Recommender and JANE, other systems used the same similarity measure for multiple subject domains which can have distinct language features. This may effect the accuracy of the results, since a similarity measure, which performs well in one subject domain, may not do so good in a different subject domain. To the best of our knowledge, Elsevier Journal Finder is the only journal recommender system that uses the same similarity measure - BM25, which is one of the five algorithms implemented in the current study. Unfortunately, Elsevier Journal Finder includes the drawbacks discussed above as it uses the same similarity measure in multiple subject domains. However, the current research minimizes this problem to some extent by testing five different algorithms for distinct subject corpora. Moreover, studies conducted by Forrester et al. (2017) and Kang et al. (2015) showed drawbacks of the Elsevier's recommender system. Accordingly, the Elsevier Journal Finder is not sufficiently capable in suggesting the correct journal for an already published article, even if the article was actually published in one of the Elsevier journals. Further, this journal finder limits search results only to the publications of the Elsevier's platform. Not extending search results at least up to the Scopus database is described as a substantial limitation of the system due to the minimum number of journal options the system compares for suggesting the appropriate ones. Moreover, its weak ranking performance for fewer corpus documents and inadequate capability of search algorithms are questioned by the two articles. IEEE Publication Recommender and JANE use only one subject domain for the recommendation process thus avoiding the use of the same similarity measure for multiple subjects. This methodology aligns with the current study since here too suitable similarity algorithms for different subjects are selected independently. Nevertheless, the current recommender system deviates from IEEE and JANE as it further expands the service by associating two subject domains and selecting them distinctly in contrast to IEEE or JANE. Also the similarity algorithms utilized in these three systems are different from each other. Using different similarity algorithms and cor-

pora are not the only differences of the new system. It has also been enhanced by associating a knowledge-based filtering component. Although, most content-based recommender systems were improved by associating a number of journal selection factors like IF level, OA provision, and author charges, those methods simply filter the results given by the content-based systems. Instead of simple filtering, in the current study the composite effect of a collection of journal selection factors is determined. This composite measure is anticipated to give more precise results than considering the factors separately. There are two common failures in the existing content-based journal recommender systems. They use only a few journal databases to select appropriate journal outlets. Further, they hardly connect with large OA journal databases. As a result, only authors who wish to publish in proprietary journals can get their benefit. Next, it is not a good practice to maintain all training documents belonging to different subject domains in a single corpus. This could also result in reduced efficiency of the recommender system. For example, there is a possibility to retrieve a medical journal title, for a social sciences input document. However, the current recommender system avoids this easily by deploying separate corpora for different subject domains.

Almost all collaborative-based recommender systems described in section 2.3 target appropriate conferences for manuscript submission. However, concepts like clustering authors with similar interests and using writing style information could also be used for recommending journals. Usually, these collaborative-based systems initiate recommendation process from content-based classification, but technically, the overall recommendation process has been much improved over the systems in section 2.2. Using composite effect of publication factors could be the main reason for these improvements. This approach is approximately similar to the concept used in the current study, but has differences due to similarity measures employed and factors unique to conferences or journals. For example, a feature like IF is so far a unique measure to journals.

Alternative journal selecting methods given in section 2.4 use logging records of jour-

nals stored in servers and separate journal selection factors. None of these approaches use content-based analysis or overall effect of journal selection factors to recommend appropriate journals.

While a number of approaches are limited to the phase of criteria development, it is noticeable that some others have geared studies towards model development and implementing recommender systems. However, we can often see that recommender systems use existing journal selection criteria for implementation tasks. The journal recommender system proposed by this dissertation uses selection factors from the previous studies as described in the next chapter. Further, it describes the methodology used for developing the new journal recommender system that can work differently than the existing systems discussed above.

Chapter 3

Methodology

"Curiosity has its own reason for existence"

– Albert Einstein: *Life Magazine* (1955), p.64

3.1 Stages of methodology

This chapter of the dissertation is organized under four major topics as follows.

1. Identifying and prioritizing journal selection factors.
2. Developing a content-based recommender component.
3. Developing a knowledge-based recommender component.
4. Configuring and evaluating the merged recommender system.

The methodology procedure of the study was also conducted parallel to the above topics and the four major stages of the development process is summarized in figure 3.1.

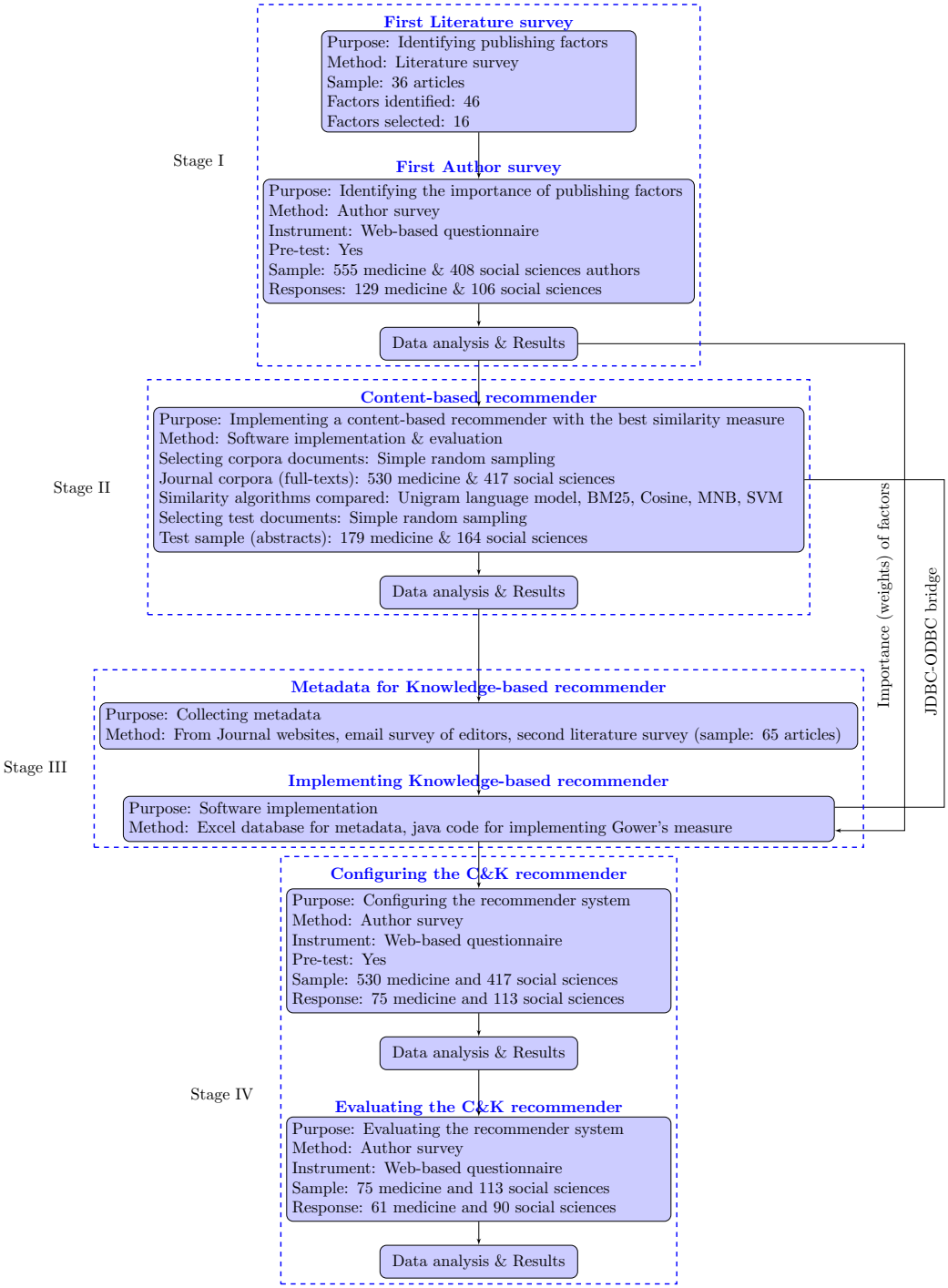


Figure 3.1: Major stages and flow of methodology

3.2 Aspects of publishing

This study was initiated with a comprehensive literature review in order to identify the factors that could influence an author's decision on selecting the most appropriate journal to submit a manuscript. 36 relevant research studies (see Appendix E) published between 1999 and 2015 were investigated to identify the various aspects of publishing (Wijewickrema and Petras, 2017). A search of computerized literature database – Google Scholar was conducted, with the following keywords: 'journal selection', 'publishing aspects', 'article submission', 'manuscript venues', and 'scholarly communication'. The possible combinations of these keywords were also used to collect relevant literature. In addition, similar studies were also identified by cross-referencing the included studies. Original research articles, conference papers, editorials, and theses were included in this collection. A consensus has been reached on the studies retained or discarded, on the basis of the following inclusion and exclusion criteria (Wijewickrema and Petras, 2017).

Study selection:

Inclusion criteria were: article submission factors, publishing in journals, article submission models. Retrieved articles with these contents were included in the study.

Exclusion criteria were: other publishing venues except journals. Retrieved articles that discuss influential publishing factors of other venues (e.g. conferences, workshops, and so on) except journals were excluded from the study.

Considering all referred articles, the study identified 46 factors (see figure 3.2) which could affect the author's decision of choosing an appropriate journal. However, on the one hand, not all these factors are recognized as important aspects of publishing by the authors of the above mentioned 36 articles. On the other hand, practically it is not feasible to deploy all these pre-identified factors in the current study. Therefore, the master list with these 46 aspects was reduced to obtain a shorter list with 16 publishing factors. The reduced list has been prepared concerning the following facts.

1. Significance of the aspect (how often they appear in literature).

2. Aspects important only for the knowledge-based analysis (e.g. factors like 'Subject coverage' are not included in the reduced list as they can be analyzed content wise).

3. Factors valid for OA journals.

4. Aspects with a quantitative value (or possible to convert to a measurable entity).

5. Aspects having a meaningful value to measure with respect to the journal (e.g. 'Colleague recommendation' does not make any sense from journal's end).

6. Factors not covered by any other factor of the list (e.g. IF and h-index are based on the same factor - citation count).

The factors excluded from the study are given below with the reasons for exclusion.

Factors included in content-analysis: Subject coverage and Previous articles on the same topic.

Factors rarely appeared in literature: Citations to same journal, Journal rank, Style and length of article, Same geographic area, Previous submissions, Article preferences, Allow to publish supplementary data, Presence in Jeffrey Beall's journal list, Publisher's contact information, Library issues, Word count, Submission deadline, Copyright issues, Type of journal, and Publication medium.

Factors not valid for OA journals: Price of journal and Access method.

Factors with no value to measure from journal's end: Clear author guidelines, Recommendation of institution, Motivation to submit, and Colleague recommendation.

Factors with no quantitative value to measure: Readership, Physical quality, Communication, Publicity, and Editorial board.

Factors covered by other factors: Number of citations received and h-index.

- Peer-review status
- Subject coverage
- Impact Factor (IF)
- Journal's reputation
- Publisher's reputation
- Abstracting and Indexing
- Time to publish
- Publication fee
- Publication frequency
- Citations to same journal
- Journal rank
- Previous articles on same topic
- Rejection rate
- Style and length of article
- International/domestic
- Same geographic area
- Representing an institution/society
- Convenience of submission
- Previous submissions
- Clear author guidelines
- Recommendation of institution
- Circulation of journal
- Readership
- Article preferences
- Availability of permanent article identifier
- Allow to publish supplementary data
- Appearance in Jeffrey Beall's journal list
- Motivation to submit
- Colleague recommendation
- Publisher's contact information
- Number of citations received
- Number of papers per year
- Physical quality
- Age of journal
- Library issues
- Price of journal
- Communication
- Word count
- Submission deadline
- Publicity
- Copyright issues
- Type of journal
- Publication medium
- h-index
- Access method
- Editorial board

Figure 3.2: Master list of factors

Then, these 16 factors were grouped again into 3 major categories, namely, Performance, Reputation, and Visibility. These primary groups can be defined as follows:

Performance: The factors that measure how well the journal executes its publication process. Most of the responsibility of deciding the magnitude/availability of each characteristic depends on the journal itself.

Reputation: Measures the factors that add a value to author's contribution and it indicates the level of appraisal received from peers (e.g. adding a value to

author's curriculum vitae).

Visibility: Measures the factors that influence the journal/article distribution or indicates their distribution.

After grouping the 16 aspects into 3 main categories, the final reduced list of factors is obtained as in the table 3.1.

Performance	Reputation	Visibility
Author charges	Impact Factor (IF)	Circulation/usage
Publishing speed	Peer-reviewed	Abstracting & indexing
Age of journal	Journal's prestige	International/domestic
Publication frequency	Publisher's prestige	Permanent article identifier
Papers per year	Presence institute/society	
Rejection rate		
Online submission/tracking		

Table 3.1: Reduced list of factors

3.3 First author survey: Manuscript submission considerations

Mere identification of the publishing factors is certainly not enough to have a good understanding about the way they affect the submission decision of the author. The importance of all these 16 factors may not be similar to each other as they measure various attributes of the publishing process. Therefore, it was imperative to determine how much importance the authors attribute to each of these 16 aspects while selecting an appropriate journal outlet. With a view to achieving this goal, first a web-based survey in the research was planned. The web-based survey methodology was specifically chosen owing to numerous advantages like higher speed, accuracy of data collection, minimum cost, and so on, over the conventional mail survey method (Fleming and Bowden, 2009; Nulty, 2008).

3.3.1 Structure of the questionnaire

As the survey instrument, a web-based questionnaire (see Appendix A) was designed with a covering email invitation to the authors of journal articles (Wijewickrema and Petras, 2017). The questionnaire consisted of 10 major questions while the first question included 16 sub-questions and the second major question with 3 sub-questions. The rest of the eight questions appeared only with their major part. The majority of the survey was based on closed-ended questions. This study used a version of the LimeSurvey[1] free and open source online survey tool customized by the Humboldt University of Berlin for the research purposes of its students and employees to design, distribute, and administrate the web-based questionnaire.

The first question with 16 sub-questions asked to rate the importance of each of the 16 publishing aspects which have already been identified in the first literature survey. This part of the questionnaire was included as a mandatory question since it is the most critical section which determines the weights given by the authors for each of the publishing aspects. Moreover, these weights can be regarded as a reflection of how importantly authors consider the given factors while selecting a journal to submit an article. A five-point Likert scale was used with 1 representing "Not important at all" and 5 representing "Very important". Many similar studies (Regazzi and Aytac, 2008; Rowlands and Nicholas, 2006; Ziobrowski and Gibler, 2000) have used the same scale ensuring the consistency and reliability of the chosen weighting system.

The second major question is an open-ended one and allowed the authors to mention and rate another three publishing concerns which do not appear under the first part of the questionnaire, yet are important in deciding an appropriate journal outlet. This component was incorporated to find out new publishing features which may be consistent with the present study, with critical influence on authors' selection criteria.

Questions three to five are included with a view to determining the author's awareness

[1] https://www.limesurvey.org/

of the existence of journal recommender systems, their usage, and to know whether they consider a journal recommender system as an important tool for selecting an appropriate journal outlet. Further, this intends to reveal the extent of impact of existing journal recommender systems and authors' general perception about their helpfulness for the works of publishing.

The sixth to ninth questions check whether there is any variation in the answers given by the respondents according to their experience of publishing, recent contributions to the field, expert knowledge of publishing process, and geographical region. Crucial tendencies of the scored ratings were expected to identify along four primary themes as described above. This is important to decide whether there is any significant impact of these issues for generalizing the authors' publishing aspects for the entire sample.

The final question is a general open-ended one to input further comments of the respondents. Here, the authors are allowed to input their independent ideas about the current study, their concerns regarding selecting journals for manuscripts submission, or any other idea relevant to the publication process.

3.3.2 Pre-test

A pre-test was conducted before sending the questionnaire to the actual participants of the survey. The main objectives of the pre-test are as follows (Thabane et al., 2010; van Teijlingen and Hundley, 2002):

1. To realize whether the respondents can understand the purpose of the questionnaire.
2. Identifying the ambiguous, difficult, and missing questions.
3. To estimate approximate time, the survey takes.

4. To check whether the adequate range of responses are provided for each question.

5. To identify the issues with logical order of the questions.

6. To know the appropriateness of general appearance of the questionnaire.

7. To study the variability of the responses and unusual tendencies of the given answers.

The questionnaire was sent to 14 participants and the responses were collected after two weeks from the date the pre-test invitation was emailed. Of the 14 targeted respondents, 13 participants completed the pre-test. They have given their comments in addition to completing the survey. The implications of the comments are given below:

It was not clear to them the reason for providing only medicine and social sciences as the respondent's subject streams to be selected. The majority of the participants emphasized the need to remove footnotes explaining the meaning of some of the 16 publishing aspects. Furthermore, the pre-test was useful to realize that some given questions were not clearly understandable and therefore should be restructured for clarity. Most respondents were not familiar with journal recommender systems and suggested explaining the concept further. In addition, there were ideas to add new questions to examine the variation of the answers over certain aspects like experience and country. Inappropriateness of categorizing questions into two sets namely, 'major questions' and 'other questions' was proposed by one respondent while pointing out the missing, but vital answer option 'neutral' for certain questions.

3.3.3 Amendments to the survey

Based on the comments given by the participants of the pre-test, the following changes were made to the questionnaire.

1. The question for stating the respondent's field of study was removed and it was determined to conduct two separate surveys with the same questionnaire for the authors of the medical domain and the social sciences field. This made the data collection process simpler and more accurate. Moreover, it helped to provide a less complicated questionnaire for the survey participants.

2. The number of footnote explanations for the questions were minimized and the questions were made self-explanatory as much as possible.

3. The logical order of the questions were rearranged so that it would help the respondents to understand the questions more easily.

4. Some questions asking about publishing factors were reworded to enhance clarity. For example, the factor 'Author charges' appeared ambiguous since there could be authors who expect no author charges. Therefore, it was changed to 'No author charges'.

5. Questions about 'Recommender systems' were unclear to the respondents. Hence, the set of questions was moved to a more appropriate place in the questionnaire (location: just after the 'Any other factors you consider' part) and started with the new question:
   "Are you aware of the existence of journal recommender systems, which can assist an author to select a suitable journal to publish (e.g. Elsevier journal finder, Edanz journal selector, Journal/Author Name Estimator, etc.)?"

6. Few new questions were added to the survey based on the pre-test comments. For example, the question, "When did you publish your first journal article?" was added in order to determine the respondent's publishing experience.

7. Some of the text boxes to input answers were replaced by drop-down lists, providing more convenience.

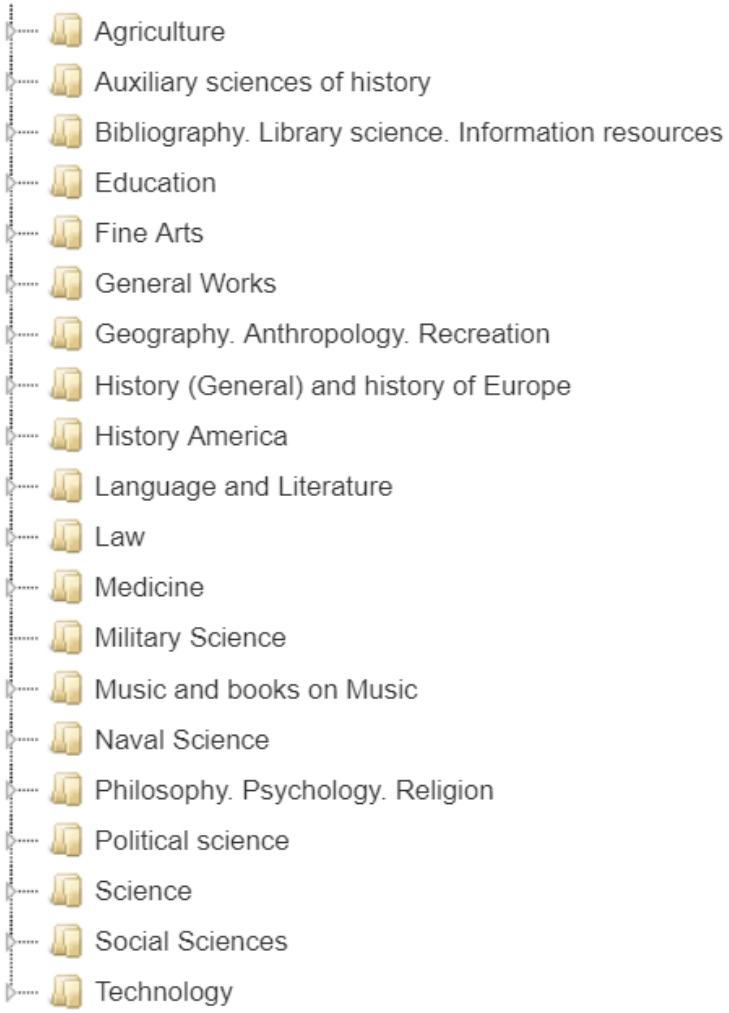


Figure 3.3: DOAJ major subject categories

3.3.4 Define populations

The aim of the current study was to select a sample of authors who have recently published their articles in OA journals. DOAJ was identified as the major source available to access OA journals. As at January 2017, it consisted of 9,485 journals representing 129 countries across the world. Thus, DOAJ was used as the source for locating journals from which to sample authors.

DOAJ has categorized its entire journal collection into 20 separate subject clusters as shown in figure 3.3. However, the population frames were restricted to the domains of medicine and social sciences since the current study considers the authors' publishing aspects from these two subject areas only. DOAJ medicine collection consisted of

journals from General medicine, Internal medicine, Specialties of internal medicine, Public aspects of medicine, Neurosciences, Biological psychiatry, Neuropsychiatry, Surgery, Neurology, Diseases of the nervous system, Dentistry, and Pharmacy & material medica while the social sciences included journals from General social sciences, Business, Commerce, Economic theory, Demography, General sociology, Industries, Land use, Labour, Management, Industrial management, Social pathology, Social & public welfare, Criminology, and Finance.

For this study, medicine and social sciences were considered as two independent populations. Therefore, authors from journals in these two domains represent the two distinct populations. The corresponding author of the most recent article published by each journal was identified as the most suitable member to represent its population. Nevertheless, the first author was selected whenever the corresponding author was not specified. If there were multiple articles published in the newest date or the publication date was not mentioned separately for each article in the latest issue, then only one author was selected randomly from the issue after arranging the authors' last names alphabetically. In most cases, the journals have organized the articles according to the date they have first appeared online as 'Electronic publication date'. Therefore, it was not difficult to pick the authors of the latest articles. In this research, the most recent articles were defined as the latest articles which were available online on DOAJ during the author's information collection period from mid December 2016 to mid January 2017. Furthermore, it was observed that those articles belonged to the journals having issues between October 2015 and January 2017.

The following facts explain the decision to limit the populations to only one (corresponding or first) author of the most recent article from each journal in DOAJ.

1. DOAJ was used as the source to obtain the two samples since the manuscript submission considerations of the authors with respect to OA journals is the focus of this study. Moreover, DOAJ is the best repository to find articles and their authors who have published in OA journals.

2. Only the authors from the most recent articles were considered because of the higher possibility of the existence of the author, validity of their contact details, author's enthusiasm to respond, possibility of remembering what they have considered when they were submitting their last manuscript, and so on. In other words, to receive a higher response rate.

3. The corresponding authors or the first authors of the article were contacted as they may be the principal investigators or research group leaders (Rowlands and Nicholas, 2006) of the particular research and were the persons mainly responsible for deciding the publication venue.

4. In order to obtain a variety of publishing aspects, authors from each journal of the considered domains of DOAJ were selected. In contrast, selecting multiple authors from the same journal could reveal similar publishing considerations. This would not help to understand the true picture of the problem.

5. Considering the availability of time and resources for the current research, the two samples were restricted to only one author and one article from each journal. We avoid selecting multiple articles from the same journal, because it could lead to reveal similar publishing concerns as stated in the previous paragraph.

There were 1,154 medical journals and 658 social sciences journals in DOAJ for the time period between mid December 2016 to mid January 2017. Accordingly, there were 1,154 authors in the medicine population while it was 658 in number for the population of social sciences (Wijewickrema and Petras, 2017). The most recent article of each of the 1,812 journals were accessed and the corresponding author's (or the first author's) full name and the email address were collected for the two subject streams separately. Then these two lists were organized according to the alphabetical order of the author's last name. Duplicate presence of the same authors were recognized and replaced by the second author of the same article. Whenever different authors with the similar last name were found, they were sorted according

to their first name.

3.3.5 Sampling and data collection

This study applied simple random sampling method (Thompson, 2012) with 95% confidence level and 3% margin of error to each population in order to obtain the samples of authors. As a result, 555 authors were drawn from the medical subject domain while 408 authors were selected from the field of social sciences (Wijewickrema and Petras, 2017). The time available to complete the study led us to make these restrictions for the sample sizes. The sample size was determined using the SurveyMonkey[2] online tool, while sample items were drawn using random tables[3].

An email invitation (see Appendix A.1) was sent to each author requesting them to take part in the web-based survey. However, since some authors' contact details were not mentioned in the articles they have published, their email addresses were not included in the population lists. In such cases, their email addresses were located from the Internet. The email message contained a hyperlink to the Humboldt University LimeSurvey database, enabling them a direct connection to the online questionnaire (see Appendix A.2). Authors who failed to respond within two weeks were sent a second email request as a reminder. However, the total response time of the survey was concluded after a week from the email reminder. It was observed that the receiving rate of responses was decreasing gradually from the starting date of the survey and become stable at the end of the third week. The survey remained live from 16 January to 06 February of 2017.

Both descriptive and inferential statistics were used in this study to analyze the data gathered from the first author survey. We applied descriptive statistics to analyze the respondent's experience of publishing. Percentage values and charts were used to represent their exposure to scholarly publishing field. For example, we used a bar-chart to compare the editorial board experience of respondents in the two subject

[2]https://www.surveymonkey.com/mp/sample-size-calculator/
[3]http://stattrek.com/statistics/randomnumber-generator.aspx

domains. Moreover, the mean value of importance given by the respondents for each journal selection factor was calculated to determine the weight of importance assigned by the authors for factors. Facts such as author's awareness and experience of using journal recommender systems were represented using percentage values. Inferential statistics was used to expose the statistically significant differences of corresponding factors between the two subject domains and to calculate the correlations between the factors. Finally, this research applied Principal Component Analysis (PCA) to understand the way these 16 journal selection factors are grouped into major categories. This directed the research to determine the average importance given by the authors for each major category. A comprehensive description of data analysis of the first author survey is included in section 4.1.

3.4 Content-based recommender system

A journal recommender system identifies similarities in the contents and the journals the most similar articles have been published in. Then a ranked list of journals based on a similarity score between the contents of the input text and journals is generated for authors to select the most appropriate journal outlet. The current chapter of the dissertation follows the same strategy to compare an input abstract described in the section namely, 'Test document sample' with a collection of already published articles described in the section namely, 'Construction of the document corpora'.

3.4.1 Introduction to tools used

Lucene

The Lucene search engine library[4] is a high performance, free, and open source Application Programming Interface (API) developed by the Jakarta project[5]. This

[4] https://lucene.apache.org/

[5] The Apache Jakarta project is a part of the Apache Software Foundation which offers a diverse set of open source Java solutions.

programme is written in Java computer programming language. Lucene works explicitly as an API, but not as an application. Therefore, it requires relatively less computational work from the user side to customize the API, since more complex programs have already been furnished.

In general, Lucene performs the following basic tasks:

1. Index building
2. Querying
3. Index searching
4. Document retrieval

As an IR tool, Lucene indexes the documents input to the system. In addition to extracting individual terms/tokens, it maintains the records of the locations of the terms along with their frequencies. Generally, Lucene builds indexes for documents in text format, but provides the option to use add-ons[6] that allow to index files in PDF, MS Word, XML, and HTML formats too (Pirro and Talia, 2007). An index of a Lucene application can be updated incrementally to minimize the time it takes for index rebuilding. Moreover, Lucene allows to form different query types including; Boolean, proximity, position-based, wildcard, fuzzy, disjunction-max, playload, and regular expression queries. Searches in Lucene can be done by giving one or more keywords. Search results are associated with a score value that indicates its similarity to the search keywords. There are BM25, BM25F, language model, information-based model, divergence from randomness similarity implementations on Lucene apart from its own VSM based algorithm (Białecki et al., 2012). In order to retrieve the documents, Lucene uses a combination of two renowned scoring methods based on VSM and Boolean model. This combination determines how relevant a document is to a user's query. Flexibility of adjusting its source code allows users to replace the

[6]Add-on is a piece of computer software which can improve the performance of another computer application, but cannot be used independently.

existing similarity algorithms by alternative algorithms and this provides a wide opportunity for researchers to test their own algorithms. The current study employs Lucene search engine library to implement the three text similarity algorithms - unigram language model, BM25, and cosine measure described in section 1.5.1.

Weka

Waikato Environment for Knowledge Analysis, or in short, Weka[7] machine learning workbench is a free and open source software tool developed by the University of Waikato, New Zealand. This tool provides provisions for data mining researchers to experiment with a bundle of already implemented algorithms and techniques for analysis. Weka is written using Java programming language and available in a few other formats besides its flexible API. Users can customize the tool by changing the API appropriately. Additionally, the code of this interface can be hacked to expose the intermediate information, which is rarely offered by other interfaces. The command line interface known as *Simple CLI* of Weka is faster than other interfaces, and demands relatively less memory for the operations. However, the user must have an adequate knowledge of various Weka commands to achieve the full benefit of this interface. Weka *Explorer* provides a graphical user interface which allows researchers to use the tool without touching the source code or executing written commands. This interface can be considered as the standard graphical user interface of Weka, which offers a simple, user friendly environment for both inputs and outputs. Weka *Experimenter* presents a comparable interface to *Explorer*, which is more suitable to conduct experiments and to run statistical tests to compare learning schemes. *Knowledge Flow* is also a graphical user interface of Weka, which affords analogous utilities to the *Explorer*. This interface utilizes graphical icons for a variety of tasks and allow them to be dragged and dropped at proper places to accomplish operations. Comparatively less memory utilization compared to *Explorer* and its capacity to learn incrementally are the key advantages of this interface.

[7] https://www.cs.waikato.ac.nz/ml/weka/

The present research employs Weka *Explorer* interface to train separate datasets as well as to classify new test documents. The *Explorer* interface enables preprocessing, classifying, clustering, learning association rules, selecting attributes, and visualizing given datasets (Witten and Frank, 2005). Ease of executing tasks is the principal benefit of this interface, in spite of huge memory consumption for larger datasets. Weka supports choosing from a wide range of acknowledged supervised classifiers for machine learning tasks. These include multiple Bayes classifiers, linear and non-linear models of SVM, k-nearest neighbour classifier, decision table, Zero R, One R, Hoffding tree, J48, random forest, random tree and many more (Frank et al., 2009). However, unlike in Lucene, all classifiers accept input data files in ARFF[8] format instead of raw text format, thus, the tool provides options to convert text files into ARFF file format.

3.4.2 System implementation

The study used the open source Lucene search engine library as the software tool to implement the three text similarity measures of the content-based recommender system. Lucene was already used in numerous studies to implement their IR systems (Gennaro et al., 2010; Hearst et al., 2007; Liang et al., 2013; Xu et al., 2008). They vary from classical text-based retrieval systems to recent image-based retrieval systems. This undoubtedly signifies the Lucene's flexibility in tuning its code for a wide variety of applications. This aspect further allows Lucene to easily incorporate indexing and searching abilities to other systems (Zhou and Xie, 2007). Therefore, Lucene search engine library was considered as the primary resource to construct the recommender system. The Lucene's standard tokenizer[9] was employed to tokenize the input text document and the corpus documents. The stop words were eliminated to remove insignificant words from the text while stemming was done to reduce the index terms to their root terms. These pre-processing steps avoid unnecessary com-

[8] Attribute-Relation File Format is an ASCII input file format developed for Weka.

[9] https://lucene.apache.org/core/6_6_0/core/org/apache/lucene/analysis/standard/StandardTokenizer.html

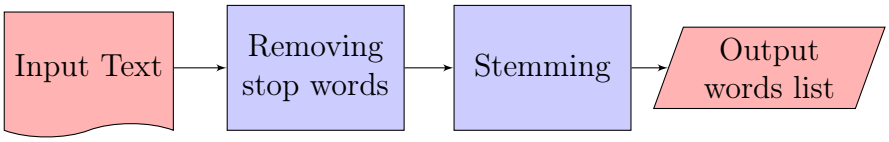


Figure 3.4: Pre-processing steps

putational work of the system. Lucene's default stop words list has been used to remove the stop words. However, the list was further improved by comparing with two other stop words lists - one from Wikimedia[10] and the other from Textfixer[11]. For stemming, the study used Porter's stemming algorithm as it is widely used for word stemming in English language (Willett, 2006; Zhou et al., 2006). In addition to Lucene, the suitability of Weka machine learning tool for implementing the content-based recommender system was assessed. The idea behind this was to either select an appropriate text similarity measure using Lucene or an appropriate supervised learning classifier using Weka tool. Weka has been recognized as a landmark system in the data mining and machine learning fields (Hall et al., 2009). Its capability of using multiple classifiers, listing classification results based on probability distributions, flexible API, and easy to understand graphical user interface led this study to consider Weka as a software tool to implement the recommender system. At the initial stage, the graphical user interface - *Explorer* was used to test the appropriateness of supervised algorithms in Weka, but it was intended to use the API at the latter stage, if the supervised classifiers perform better than similarity measures. Besides, the same stop words list and the stemming algorithm used in Lucene were also implemented in Weka for data pre-processing. Also, Weka's filter called `'StringToWordVector'` was used to convert string attributes of the text into a set of numeric attribute as Weka does not handle strings directly. Figure 3.4 illustrates the order of pre-processing steps.

[10] http://meta.wikimedia.org/wiki/MySQL_4.0.20_stop_word_list
[11] http://www.textfixer.com/resources/common-english-words.txt

Construction of the training document corpora

Since the current study is limited to addressing the issues of only two subject domains, we confined the acquisition of corpus documents to the medicine and social sciences domains. Two separate training document corpora were built for these domains. Already published full-text articles in OA journals were selected as appropriate training documents to include in the corpora. Full-text articles were selected, because they include more information than a short piece of text such as an abstract. This approach is likely to generate more accurate recommendations than using a corpus with short pieces of text. We identified the DOAJ as the suitable resource to locate OA journals. DOAJ usually lists journals based on their subject categories and this facilitated the extraction of medicine and social sciences journals separately. However, all subject relevant journals listed in the DOAJ in the corpora were not included. In general, DOAJ consists of several OA journals published in non-English languages like Portuguese, Spanish, German, Arabic and Turkish. These journals were not considered for inclusion since it was intended to build the recommender system only for submissions in English language. Multilingual journals, with English language articles, were also considered as valid journals to include in the training corpora as they allow submissions of manuscripts in English. There were 530 relevant medicine journals and 417 relevant social sciences journals in DOAJ during the corpus documents acquisition period from early-May 2017 to mid-July 2017 (see Appendix G) (Wijewickrema et al., 2019). Articles published in the most recent year available during the document acquisition period were selected to include in the corpora. It was observed that the number of issues per year published by different journals vary from one journal to another and the number of articles depends on this fact. However, on average, most of the journals publish at least 10 articles per year. Therefore, it was decided to obtain the 10 newest full text articles from each of the social sciences and medicine journals to build the corpora. When there were more than 10 articles published within the year, the corpus collection was limited to the 10 most recent articles. If there were less than 10 articles within the considered year,

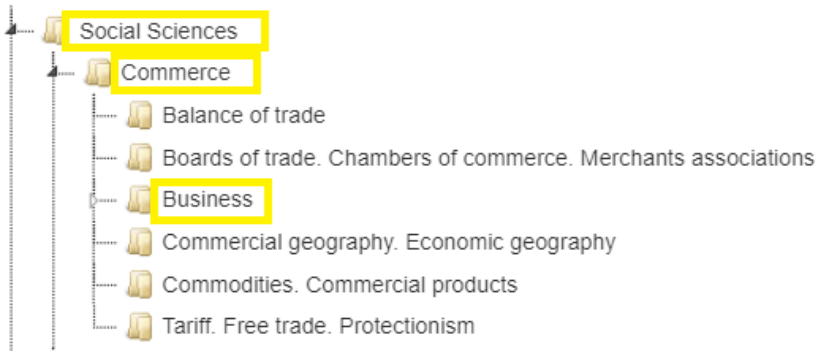


Figure 3.5: Example for subject breakdown

the rest were included from the latest articles of the previous year. Also, the articles from special issues were not included in the training corpora as they could consider only a specialized facet of the journal. All full-text articles were converted into text format to store in the training corpora after they were downloaded manually from the official websites of each of the journals. After collecting all corpora documents, there were 5300 training documents in the medicine corpus while the social sciences corpus included 4170 training documents (Wijewickrema et al., 2019). All journals listed in the DOAJ are classified by its editorial staff according to the Library of Congress Classification (LCC)[12]. Usually, the classification is assigned to a journal when it is registered with the DOAJ service. Therefore, we considered these classifications as the academic sub-disciplines of the journals since the process is systematic and controlled. It can be observed that some of the journals in this collection have been categorized down-to very specific subjects considering their nature, while others are not. Specific journals are classified into two or three levels down from their broadest subject level.

e.g. Social sciences: Commerce: Business

Figure 3.5 illustrates that the considered journal belongs to 'Business' which is three subject levels down from 'Social sciences'. The most specific given sub-discipline is considered as the relevant subject category for each journal. Also, there were some journals belonging to more than one sub-discipline category. A number of studies (McKiernan et al., 2016; Pong et al., 2007; Prabowo et al., 2002; Yong Wang and

[12]https://www.loc.gov/catdir/cpso/lcco/

Tang, 2003) have already used LCC scheme as the key classification framework to establish and evaluate IR systems and they provide further evidence to acknowledge LCC as an appropriate resource to classify journals into their sub-disciplines.

3.4.3 Text similarity measures and classifiers

Correct identification of an appropriate text similarity measure or supervised learning classifier for comparing an input document and the corpus documents is a crucial point of this study. Previous studies (Bogers and van den Bosch, 2007; Clarke et al., 2002) reveal that the best similarity measure could depend on the nature of the problem someone investigates. Islam and Inkpen (2008) emphasize the domain dependence of different text similarity measures. Nature of the vocabulary used in independent subjects, number of corpus documents, lengths of the corpus documents, and structured or unstructured nature of the texts may decide the most appropriate text similarity measure or supervised learning classifier for a recommender system. Therefore, three major string-based similarity measures and two supervised learning classifiers were implemented and evaluated separately to identify an appropriate one for the current problem. The number of similarity measures and learning classifiers which were evaluated was restricted to five as it was not practical to test more algorithms, considering the time factor to complete the research. However, these three similarity measures were selected from three distinct IR models that could reflect higher diversity in their performance. Also, two supervised learning classifiers were selected from different classifier families expecting more diversity in the results.

The unigram language, BM25, and cosine similarity measures were implemented in the recommender system in order to determine the most appropriate one for each of the subject domains medicine and social sciences, while SVM and Multinomial Naïve Bayes (MNB) were chosen to implement as supervised learning classifiers. The details of these similarity measures and learning classifiers are included in the next few paragraphs.

Similarity measures

The unigram language measure is computed using equation (3.1) (Zhai, 2008)

$$p(t_1, t_2, t_3, \ldots, t_n) \approx \prod_{i=1}^{n} p(t_i), \tag{3.1}$$

where

$$p(t_i) = \frac{tf_{d,t_i} + 1}{\sum_{j=1}^{m} tf_{d,t_j} + v}. \tag{3.2}$$

We applied 'Add-One smoothing' (Laplace smoothing) in order to avoid zero probabilities when the query terms do not appear in the corpus documents (Schütze et al., 2008; Zhai and Lafferty, 2004). Here, m, n, v, and tf_{d,t_j} represent the total number of terms in the given corpus document d, the total number of terms in the query document q, the number of distinct terms (vocabulary) in the corpus document, and the frequency of term t_j in document d, respectively.

BM25 measure is defined by equation (3.3) (Sixto et al., 2016)

$$\text{score}(d, q) = \sum_{i=1}^{n} idf_{t_i} \frac{(k_1 + 1) tf_{d,t_i}}{k_1\left(1 - b + \frac{b|D|}{avgdl}\right) + tf_{d,t_i}}, \tag{3.3}$$

where $|D|$ denotes the total number of terms in the document d, while $avgdl$ denotes the average number of terms in a corpus document. The values of the two free parameters, k_1 and b are set to 1.2 and 0.75 respectively, as it was experimentally proven by Jones et al. (2000) to give the best retrieval performance for text corpora as used by the current study. Moreover, these are the default tuning parameters used in the Okapi system (He and Ounis, 2003). idf_{t_i} gives the inverse document frequency of each query term, which is defined as in equation (3.4) (Apache Software Foundation, 2017),

$$idf_{t_i} = \log\left(1 + \frac{N - n_{t_i} + 0.5}{n_{t_i} + 0.5}\right), \tag{3.4}$$

where N is the total number of documents in the corpus and n_{t_i} denotes the number of documents containing the query term t_i.

Equation (3.5) shows the cosine similarity (Adomavicius and Tuzhilin, 2005; Baeza-Yates and Ribeiro-Neto, 1999).

$$\text{score}(d,q) = \frac{\sum_{i=1}^{n}(\overline{tf_{q,t_i}} \cdot idf_{t_i})(\overline{tf_{d,t_i}} \cdot idf_{t_i})}{\sqrt{\sum_{i=1}^{n}(\overline{tf_{q,t_i}} \cdot idf_{t_i})^2}\sqrt{\sum_{j=1}^{m}(\overline{tf_{d,t_j}} \cdot idf_{t_j})^2}}, \tag{3.5}$$

where $\overline{tf_{q,t_i}}$ and $\overline{tf_{d,t_i}}$ denote the normalized term frequencies of the term t_i in query q and document d respectively. They are defined as follows:

$$\overline{tf_{q,t_i}} = \frac{tf_{q,t_i}}{\sum_{i=1}^{n} tf_{q,t_i}}$$

and

$$\overline{tf_{d,t_i}} = \frac{tf_{d,t_i}}{\sum_{i=1}^{m} tf_{d,t_i}}.$$

The term idf_{t_i} in the equation 3.5 is the inverse document frequency of the term t_i, defined by equation (3.6) (Apache Software Foundation, 2010),

$$idf_{t_i} = 1 + \log\left(\frac{N}{n_{t_i}+1}\right). \tag{3.6}$$

According to the equations (3.4) and (3.6), it is clear that the two measures: BM25 and cosine similarity, use different inverse document frequency terms for calculations. This is one of the primary differences between these two algorithms.

Supervised learning classifiers

The Multinomial Naïve Bayes algorithm complies with the basic assumptions of the Naïve Bayes model, yet it employs the multinomial distribution as its probability distribution. MNB uses Bayes rule to compute the highest probability of a given query document q to be classified under category c. MNB classifier is obtained as follows starting from the Bayes rule (Kibriya et al., 2004).

$$p(c|q) = \frac{p(c)p(q|c)}{p(q)}, \qquad c \in C \tag{3.7}$$

where the set of categories is denoted by C and $p(c)$ is estimated by dividing the number of corpus documents in category c by the total number of corpus documents. Further, $p(q|c)$ and $p(q)$ are defined as:

$$p(q|c) = \left(\sum_{i=1}^{n} tf_{q,t_i}\right)! \prod_{i=1}^{n} \frac{p(t_i|c)^{tf_{q,t_i}}}{tf_{q,t_i}!}, \tag{3.8}$$

and

$$p(q) = \sum_{j=1}^{|C|} p(j)p(q|j). \tag{3.9}$$

Add-one smoothing as defined by equation (3.2) is used in the following form to estimate the term $p(t_i|c)$ in equation (3.8):

$$p(t_i|c) = \frac{tf_{c,t_i} + 1}{\sum_{j=1}^{V} tf_{c,t_j} + v},$$

where tf_{c,t_i} denotes the count of the term t_i in the all corpus documents belonging to class c. A common assumption of the Naïve Bayes models is the strong independence among the document features, which are utilized in classifying query documents.

Support Vector Machine is a powerful supervised learning method frequently used in the problems of classification and regression analysis. SVM builds appropriate models to assign categories for given test documents using already labeled training documents in a corpus. A model of an SVM can be illustrated as a set of points in space, representing labeled training documents, which are divided into distinct categories with clear margins. The mechanism allows mapping new documents to points in the space thus enabling the prediction of their matching categories. Basically, SVM attempts to find the hyperplane (e.g. a line in two dimension), which separates the nearest points of two categories while maintaining the maximum gap between them. For example, the line l_2 is a better separation line than the line l_1 for the two categories c and c' in figure 3.6 as it maintains the maximum gap between the two points p and p', in the two categories.

However, the optimal separation hyperplane of two categories may not necessarily

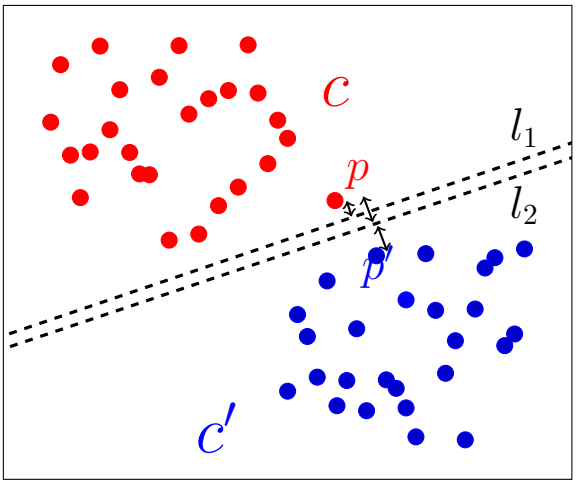


Figure 3.6: Optimal separation hyperplane between two groups of data points

be a simple linear one always. There can be more complex non-linear hyperplanes that optimize the separations of nearest points belonging to distinct categories. Consequently, Cortes and Vapnik (1995) suggest solving the dual optimization problem given by equation (3.10) to train a soft margin (or non-linear) SVM classifier using maximum-margin hyperplanes.

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j), \tag{3.10}$$

$$\text{under constraints: } 0 \leq \alpha_i \leq p, \qquad i = 1, 2, \ldots, N,$$

$$\sum_{i=1}^{N} y_i \alpha_i = 0,$$

where N, x_i, y_i, α_i, p, and $W(\alpha_i)$ denote the number of data points (size of training corpus), ith data point, binary category corresponding to x_i, Lagrange multipliers (Lagrange, 1853), SVM hyperparameter, and quadratic function of α_i respectively.

A variety of methods are proposed in literature to train SVMs. Amongst others, Sequential Minimal Optimization (SMO) introduced by Platt (1999a) is significant due to its simplicity and computational cost effectiveness compared with the other existing methodologies. SMO functions as an iterative algorithm and decomposes

the original problem into the simplest possible sub-problems. These sub-problems are analytically solvable and it appears to be the greatest advantage of SMO, as this avoids using complex numerical optimization procedures. The current study uses SMO method as the training algorithm in the SVM classification component.

Score for ranking journals

The recommender system computed a relevance score for the input test documents described in the section namely, 'Test document sample' and each journal in the corpus. The journal that obtains the highest relevance score was listed as the top rank journal while the journal with the lowest relevance score was listed as the bottom rank journal. Following this method, an ordered list of journals was generated based on the journals' relevance to the input test document. Usually, text similarity measures compute similarity scores between the input test document and each training document in the relevant corpus. Recall that each corpus journal includes 10 training documents. The average of the total score accounting for all 10 training documents belonging to each journal was computed for obtaining an average similarity score between the test document and the corpus journals (Wijewickrema et al., 2019). This procedure led to obtain a single score between an input test document and each journal in the considered training corpus. Finally, these average scores were organized by the journal titles starting from the highest score to the lowest. However, since supervised learning classifiers usually do not generate similarity scores, appropriate probability distribution of input document in each corpus journal was fitted as the measure to obtain an ordered list for journal's appropriateness. Here, the direct probability distribution scores given by the MNB were used, but an informative probability distribution output for SVM failed to be obtained as it usually generates binary probability distribution scores. Therefore, the method introduced by Platt (1999a) for mapping SVM results to logistic models was used to obtain continuous probability distributions to rank journals. Then, similar to text similarity measures, these probability distributions were organized by the journal titles starting from the

highest score to the lowest to generate the desired list.

Although, there was no specific configuration to be adjusted for the MNB, SVM allowed to change its configuration. However, this study used the default configuration parameters in table 3.2 for SVM since they could be valid for most cases. The

Configuration parameter	Parameter value
Classifier	Sequential Minimal Optimization (SMO)
Epsilon	1.0E-12
Kernel	Polykernel E1.0 C250007
Complexity parameter	1.0

Table 3.2: Configuration parameters of SVM

study selected SMO as the SVM classifier of the current problem. A description of SMO and the reasons for selecting SMO as the SVM classifier are given in the section for 'Supervised learning classifiers'. The kernel function of SVM works like a similarity function to determine the similarity between two vectors. This approach supports to compare test and training documents. The default SMO kernel function of Weka – the polynomial kernel was applied by the current study. It has given better results than other kernel functions in some studies (Nanda et al., 2018). The polynomial kernel represents the similarities of vectors using polynomials of original variables. The epsilon parameter of SMO represents the exponent of the kernel function. The default exponent 1 is used by this study since the non-linear kernels with degree greater than 1 take higher computational time for training than linear kernels take (Platt, 1999b). The complexity parameter is used to build the hyperplane between two classes. More information about hyperplane can be found in the section namely, 'Supervised learning classifiers'. This parameter can be adjusted appropriately to avoid misclassification rate. Usually, large values of the complexity parameter minimize the misclassification rate.

3.4.4 Evaluation of content-based recommender system

Test document sample

The study collected appropriate samples of test documents to evaluate the content-based recommender system. Moreover, DOAJ was selected as the source to locate all these test documents. There were two major reasons to collect test samples from DOAJ. The possibility of classifying test documents based on the corresponding subject categories which have already been assigned to journals in DOAJ was one of the benefits. In addition, selecting pre-classified test samples from elsewhere (e.g. from Scopus) could lead to incompatibilities between the methods followed to classify test and training documents.

The simple random sampling method was applied separately to the medicine and social sciences journal populations to select two samples of journals to extract test documents. 90% confidence interval and 5% margin of error were allowed to collect 179 medicine and 164 social sciences journals after preparing the journal title lists according to the English alphabetical order (Wijewickrema et al., 2019). We applied a different criteria to select the test document sample than the criteria – 95% confidence interval and 3% margin of error, which was utilized for the first author survey (see section 3.3.5). The current sample selection criteria was applied in order to reduce the number of test documents since it gives an adequate sample size for testing. Furthermore, the limited time availability to complete the evaluation instigated adjusting the selection criteria as above. The same online calculator tools described in section 3.3.5 were used to determine the sample sizes and to generate random tables for selecting sample entities. Thereafter, the most recent abstract (along with title and keywords) was selected as the test document from each journal of the two samples. However, since we have already included the most recently published 10 articles in our training set, we had to ensure the absence of the selected abstract in the training set to make the test entities unseen. Whenever the selected abstract for testing was already included in the corpora, we ignored the first 10 recently pub-

lished articles and selected the next one from the same journal. Only abstracts were included in the test document sample, because it is unlikely an author feeding the entire content of a fresh article to a recommender system for assistance. The test documents were assigned to different sub-disciplines based on the corresponding sub-disciplines of the journals in which they have been published. Accordingly, there were 179 test documents of the medicine domain assigned to 38 distinct sub-disciplines, while there were 164 test documents of the social sciences domain assigned to 31 distinct sub-disciplines. Figures 3.7 and 3.8 show the distribution of test documents by sub-disciplines.

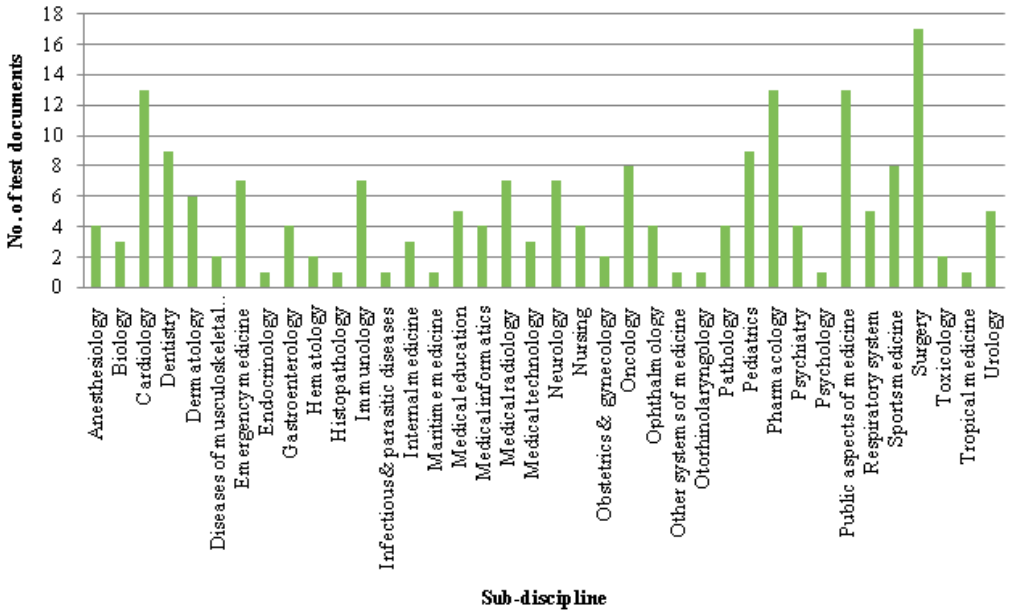


Figure 3.7: Number of test documents belonging to sub-disciplines of the medicine

Evaluation criteria

This section is dedicated to developing a suitable guideline to evaluate the retrieved results. In general, one can consider the relevance of retrieved results based on one of the following two ideals:

1. Relevance judgment is binary for the retrieved results (i.e. the result is relevant

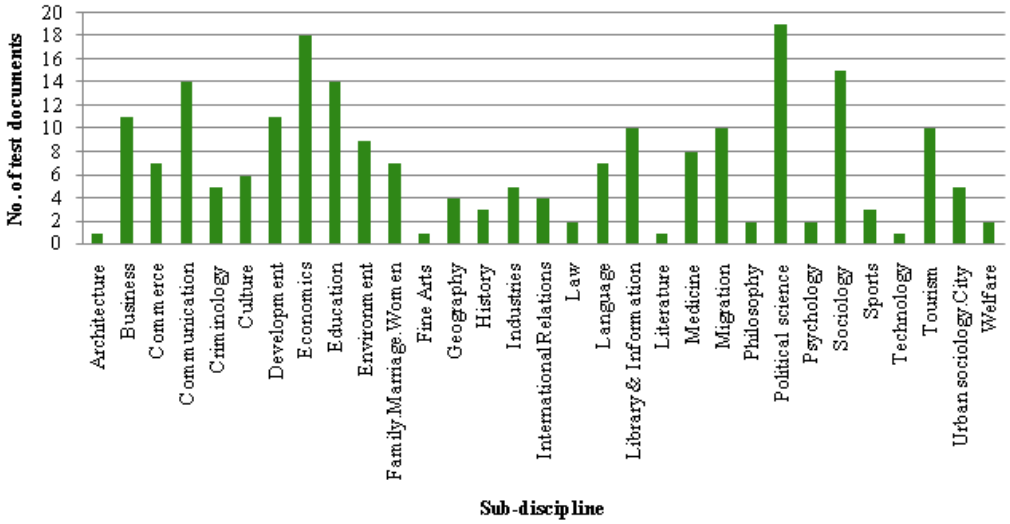


Figure 3.8: Number of test documents belonging to sub-disciplines of the social sciences

or non-relevant)

2. Relevance judgment is graded/weighted for the retrieved results (i.e. the result has intermediate values like, partially relevant).

The simplest method, binary judgment is the most commonly applied method for general IR cases. However, the second method has been used by a number of studies for web search engine evaluation (Chu and Rosenthal, 1996; Clarke and Willett, 1997; Ding and Marchionini, 1996; Shafi and Rather, 2005; Usmani et al., 2012). For example, Chu and Rosenthal (1996) have calculated precision scores based on a three-level (1 for relevant, 0.5 for somewhat relevant, and 0 for irrelevant) scoring criteria. In contrast, Usmani et al. (2012) used a five-level criterion to determine the precision scores. Considering both the nature of subjects of corpora journals and the hierarchical order of subjects' arrangement in LCC, we adopted the relevance on multiple levels to evaluate the current problem. We used the sequential order of sub-disciplines as they appear in the LCC to define sub-discipline levels and corresponding magnitudes (i.e. 1, 2, 3, 4, and so on, see figure 3.9). These magnitudes were used as the weights of relevance between an input abstract document and the retrieved journal. Moreover, the following criteria were used to make the relevance

judgment (Wijewickrema et al., 2019):

Relevant case: The sub-discipline of the retrieved journal is a perfect match with the sub-discipline category of the input abstract or the journal belongs to more specific sub-discipline category than the input abstract, yet a broader sub-discipline category of the journal equals to the sub-discipline category of the abstract at the same sub-discipline level of the abstract. In this case, the magnitude of the sub-discipline level of the abstract is assigned as the weight of relevance.

Example i: When the retrieved journal belongs to 'Groups and organizations' or 'Community' while the input abstract belongs to 'Groups and organizations', we assign a weight of 3 because the sub-discipline level of the abstract is 3 (see figure 3.9). Moreover, 'Groups and organizations' is a broader sub-discipline of the sub-discipline 'Community' of the retrieved journal and tallies with the sub-discipline of the abstract at the sub-discipline level 3.

Example ii: When the retrieved journal belongs to 'Sociology' or any other sub-discipline of 'Sociology' (i.e. 'Culture', 'Groups and organizations', or 'Community') while the input abstract belongs to 'Sociology', we assign a weight of 2 because the sub-discipline level of the abstract is 2 (see figure 3.9). Moreover, 'Sociology' is a broader sub-discipline of the sub-disciplines 'Culture', 'Groups and organizations', or 'Community' of the retrieved journal and tallies with the sub-discipline of the abstract at the sub-discipline level 2.

Less relevant case: The sub-discipline of the retrieved journal is broader than the sub-discipline category of the input abstract, yet a broader sub-discipline category of the abstract equals to the sub-discipline category of the journal at the same sub-discipline level of the journal. In this case, the magnitude of the sub-discipline level of the journal is assigned as the weight of relevance.

Example i: When the retrieved journal belongs to 'Sociology' and the input abstract belongs to 'Culture', we assign a weight of 2 since the sub-discipline level of the journal is 2 (see figure 3.9). 'Sociology' is a broader sub-discipline of the sub-discipline category 'Culture' of the input abstract and it aligns with the sub-discipline of the journal at level 2.

Example ii: When the retrieved journal belongs to 'Social sciences' and the input abstract belongs to any other sub-discipline of 'Social sciences' (i.e. 'Commerce', 'Sociology', 'Business', 'Culture', 'Groups and organizations', or 'Community'), we assign a weight of 1 since the sub-discipline level of the journal is 1 (see figure 3.9). 'Social sciences' is a broader sub-discipline of the sub-discipline categories 'Commerce', 'Sociology', 'Business', 'Culture', 'Groups and organizations', or 'Community' of the input abstract and they align with the sub-discipline of the journal at level 1.

Irrelevant case: The retrieved journal falls into none of the cases given by 'Relevant' or 'Less relevant'. In this case, 0 is assigned as the weight of relevance.

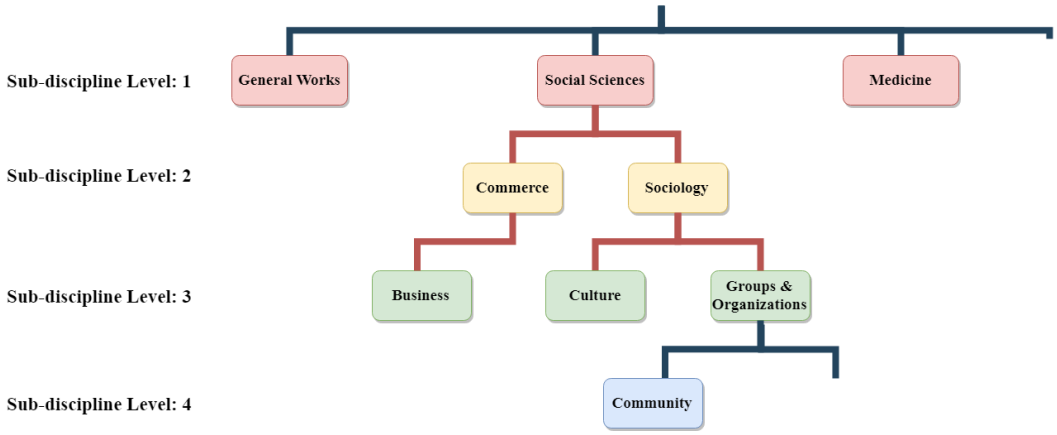


Figure 3.9: LCC hierarchy of sub-disciplines arrangement

Evaluation metric

We considered relevance of the retrieved results as the measure which reflects the performance of five algorithms under study. IR systems evaluation studies have used

numerous evaluation metrics to assess performance of their systems. Precision and recall are regarded as basic, but widely used classical retrieval performance measures available at present (Kumar and Gupta, 2015). $F-$measure is a combination of precision and recall which is also commonly applied for binary judgment tasks (Busa-Fekete et al., 2015). In addition to these, there exist several relatively advanced measures like Mean Average Precision (MAP), Receiver Operating Characteristics (ROC) curve, Expected Reciprocal Rank (ERR), and Mean Average generalized Precision (MAgP) (Chapelle et al., 2009; Kamps et al., 2007). However, this study applied the Normalized Discounted Cumulative Gain (NDCG) measure introduced by Järvelin and Kekäläinen (2002), since the relevance judgment was based on a weighted scheme. Usually, NDCG allows to assess the retrieved results of an IR system based on both weighted relevance judgments and ranking. Also, NDCG can be considered as a widely used evaluation metric across a diverse range of IR applications (Sanchez Bocanegra et al., 2017; Busa-Fekete et al., 2012). Therefore, we considered NDCG as an appropriate metric to evaluate the results of this study. The formula of the NDCG used by the study is given below.

Equation (3.11) (Järvelin and Kekäläinen, 2002) computes the NDCG of the first p retrieval results for a given query.

$$NDCG_p = \frac{DCG_p}{IDCG_p}, \tag{3.11}$$

where DCG_p is the Discounted Cumulative Gain at rank position p, defined as (Carterette and Jones, 2008):

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2(i+1)}. \tag{3.12}$$

Here, rel_i is the graded relevance of the retrieved result at rank position i.

$IDCG_p$ is defined as the Ideal Discounted Cumulative Gain at rank position p:

$$IDCG_p = rel_1 + \sum_{i=2}^{|REL|} \frac{rel_i}{log_2(i+1)}, \tag{3.13}$$

where, $|REL|$ indicates the list of relevant documents (up to rank position p) ordered by their relevance. All of the documents in two journal corpora and test sample were classified according to LCC scheme. Therefore, the information about sub-disciplines of each journal in two corpora and the test documents were easily adapted to determine the IDCG values.

A considerable number of IR systems evaluation studies have limited their evaluations to the first 10 retrieved results (Chu and Rosenthal, 1996; Vaughan, 2004). Little interest of the user with the later part of retrieved results could be one of the reasons behind this selection. This condition makes more sense for recommender systems developed for journals. In general, an author has no reason to consider too many retrieval results to select a suitable journal outlet to make a submission. Journals spend relatively higher time span for review process and to reach a decision. Hence, practically it is worthless to consider too many lesser relevant journals as one cannot make simultaneous submissions for multiple journals. A research done by Silverstein et al. (1999) for web search engine evaluation further supports this argument. Considering all these, we have used only the first 10 retrieved results for performance evaluation.

3.5 Knowledge-based recommender system

3.5.1 Journal metadata

Section 3.2 of this research has investigated 16 factors that could influence the authors' choice of deciding an appropriate journal outlet for submitting their manuscripts. This part of the dissertation describes these factors in more detail with the collection of corresponding data for these 16 factors concerning each journal in the two subject corpora. Hence, it allows to retrieve journal titles via combining authors' priority criteria of journal selection with relevant data available for each journal in the corpora.

Representing a society or an institution: The journals that represent a professional institution, society, or an association could gain more attention of authors than other journals. Regazzi and Aytac (2008) argue that this factor was considered very positively by the participants of their focus group discussions.

Author charge: Some journals charge an author fee to publish an accepted article. Processing cost of a manuscript, payments for editorial staff, and the cost for printing could be covered by journals from the author fee they receive (Welch, 2012). Author fee may influence more on the OA authors as they tend to charge an author fee to provide the OA facility (Shokraneh et al., 2012). However, a higher author fee is likely to increase the distance between a journal and the author. Some journals charge only submission fee or APC as the author charge, while others charge both of them to process an article. The present research consider including one of them if the journal charges only one fee and both of them when the journal charges both, as the author fee of a journal.

Availability of a permanent article identifier: A permanent article identifier is a unique identification of an electronic document and do not change over the lifetime of the document. Therefore, a permanent article identifier provides more stable and reliable identification of a document than a Uniform Resource Locator (URL). For example, Digital Object Identifier (DOI) is a renowned permanent article identifier which helps to locate an electronic document. The journal's ability of providing a permanent article identifier for an article is an added advantage for the authors. There are different types of permanent article identifiers including Archival Resource Keys (ARKs), Uniform Resource Names (URNs), DOIs, Extensible Resource Identifiers (XRIs), and so on. However, the presence of any of these permanent article identifier is defined as the 'Availability of a permanent article identifier' and included as a journal selection factor for the current study.

Age of journal: Journals can improve the quality, reputation, readership, circulation and many other publishing aspects with the experience they gain. More-

over, the age of a journal can be the most important determinant of its experience. Age of a journal is usually counted from the date it initiated. There can be journals started with a different name to the name they have at present. The current study considers a journal's age from the date it initiated with the name it uses at present.

Peer reviewing: Peer review process allows evaluating a submitted manuscript by a number of experts in the relevant field. The number of reviewers depends on the available resources and review policy of the journal. This process reviews the contents of a manuscript based on a number of facts such as relevance, accuracy, significance, originality, and quality. The peer review process can take many forms such as single-blind, double blind, post-publication, open, and so on. However, we accounted the presence of any of these peer reviewing types for the current study.

Publishing speed: Long time span that takes to publish an article from its initial submission is one of the major issues associated with the publication process. Time for editor's initial inspection, external reviewing, author corrections, and final editing are the basic time slots of the publication process. It is obvious that a long publication time span discourages the authors to submit their manuscripts for some journals. Time (in weeks) from the first submission to the first online appearance of an article is considered by the current study to assess this factor.

Impact factor: For any given year, the IF of a journal is the average number of citations received by an article published in that journal during the two preceding years (Garfield, 1972). For example, the IF for year 3 is obtained by dividing the number of citations received in years 1 and 2 by the articles published in these two years by the total number of items published by the journal in years 1 and 2.

$$\text{IF in year 3} = \frac{\text{Number of citations for items published in years 1 \& 2}}{\text{Total number of items published in years 1 \& 2}}$$

Journal's prestige/reputation: Reputation of a journal can be described as the opinion about the journal made by scholars in a discipline relevant to the journal. This opinion is built upon the concerns like community trust, usage, popularity among the scholars, and so on. A prestigious journal in a discipline can be selected as a good option for publishing since it could be impossible to exist a journal with all the best aspects a journal outlet must attain (Klingner et al., 2005). Knight and Steinbach (2008) describe prestige and credibility as factors that based on perception. According to González-Pereira et al. (2010), SJR value of a journal can be considered as a measure of the prestige factor.

Publisher's prestige/reputation: Publisher of a journal can make a huge impact on author's decision. Author would like to publish their articles under a reputed publisher due to higher reward they receive. However, the reputation of a publisher is highly likely to depend on the journals published. High reputation achieved by the journals published could also influence the reputation of the publisher. We used this argument to include the factor – 'publisher's reputation' and measured it using the average SJR of the journals belonging to the considered publisher.

Online submission with tracking facility: Most of the present journals facilitate authors to submit articles online. These submission systems also provide tracking facilities to monitor the processing stage of a manuscript at any time. Özçakar et al. (2012) reveal that approximately 70% of authors in their sample concerned this factor as an important factor of the journal selection procedure.

Acceptance rate: The acceptance rate of a journal can be defined as the number of accepted articles divided by the number of submitted articles per year/issue. Reasons for rejecting articles for publication can vary from one journal to another. Further, this measure can be considered as a quality control benchmark of a journal since it depends on positive reviews, journal's policy, and the opinion of the journal's editorial board.

International/domestic: Journals can be categorized into two parts namely, domestic and international. There can be domestic journals which do not accept articles from the scholars of other countries. Usually, the international journals are published in English language, but may also allow submitting in few other selected languages. Author contributions received by a journal from different countries is an indicator of the journal's international exposure. Average percentage of articles per issue authored by authors outside the country of origin of the considered journal is defined to measure this factor in this study.

Number of papers published per year: Journals may have their own reasons to restrict the number of papers per year. Minimizing expenditure, avoiding unnecessary workload, and selecting only high quality articles can be considered as some of the major reasons behind these restrictions. This study considered all types of full-text articles published by a journal per year to measure this factor.

Publication frequency: Usually, journal issues are published periodically. The number of issues published per year is generally defined as the frequency of a journal. This frequency may vary from one journal to another as annually, biannually, quarterly, monthly, and so on. It is obvious that the possibility of publishing more articles increases with the number of issues per year. This opens a wide opportunity for authors to publish their articles.

Abstracting and Indexing (A&I) services: A journal's inclusion in A&I services assist to improve the visibility of an article published in the journal. Usually, A&I services evaluate the performance of a journal to include the journal in their database. Therefore, the authors could measure the quality of a journal based on the criteria used by the A&I service to include the journal in their list. Several A&I services are available at present to evaluate journals to include in their databases. Some of these services and the criteria the current research used to rank their influence on medicine and the social sciences based on how often these services are discussed in the literature are described in the

next section.

Circulation or usage: Annual usage or circulation of a journal can be determined based on the number of subscribers a journal collected within a year. Higher circulation of a journal is substantial to promote the journal among scholars. Wide circulation of a journal is important for the authors, journals, and publishers too. Wide circulation is likely to increase the number of citations for the articles published, while the authors get confident to submit more articles as their personal citation counts are increased. Number of full-text article downloads per year is used by this study to measure the factor.

Table 3.3 shows the primary source from where we have collected data for each factor. In addition, it explains the method for measuring each factor. These data are based on the year 2016 as they were the latest statistics available during the period of data collection.

Whenever the relevant data were not presented in the primary source, we contacted the editor (see Appendix B) of the corresponding journal to obtain required information. All journal metadata were stored in a Microsoft Excel file.

Evaluating abstracting and indexing services

A collection of previously published articles were studied to identify the most influential A&I databases for social sciences and medicine journals. The second literature survey collected 65 articles (see Appendix F) from Google Scholar. The search was limited to articles published between 2000 and 2018 since the electronic A&I databases were not widely used in the years before then. Moreover, the articles were limited to the year 2018 as the current part of the research was conducted in 2018. The following search strings were used to retrieve relevant articles from Google Scholar.

'Journal abstracting and indexing'

Factor	Primary source	Numerical measure
Representing society/institution	DOAJ	0 or 1[a]
No author charges	DOAJ	0 or 1[b]
Permanent identifier	DOAJ	0 or 1
Age	Journal's official website	Age in years
Peer review status	DOAJ	0 or 1
Time to publish	DOAJ	Number of weeks
Impact factor	Journal Citation Reports	IF value
Journal prestige	SCImago official website	SJR[c]
Publisher prestige	SCImago official website	Average SJR[d]
Online submission with tracking	Journal's official website	0 or 1
Acceptance rate	Journal's official website	Percentage
Authors from different countries	Journal's official website	Percentage[e]
Number of papers per year	Journal's official website	Number
Number of issues per year	Journal's official website	Number
Abstracting/ Indexing services	Journal's official website	Number[f]
Number of subscribers	Journal's official website	Number[g]

[a] 0 for absence and 1 for presence.
[b] Amount in USD when 0.
[c] We considered SJR value as an appropriate indicator to measure the prestigious factor of journals (González-Pereira et al., 2010).
[d] Calculated the average SJR value for all journals published by the same publisher.
[e] Calculated the average percentage of international authors per issue.
[f] Explained in table 3.4
[g] We used the number of downloads per article to represent the number of subscribers. All OA journals are available online, but print versions are not available for most of them. Therefore, considering the subscribers only for print version limits utilizing the factor for journals published in both online and print versions. Moreover, editors and publishers may not be enthusiastic about collecting circulation statistics for OA journals as they are free to access.

Table 3.3: Metadata types, sources and numerical forms

'Abstracting and indexing databases social sciences'

'Abstracting and indexing databases medicine'

The retrieved results were sorted by their relevance to the search string. Also, only the first 100 results from each search string were considered for the study. As a result, 300 search results were analyzed to select an appropriate sample of previous studies. The presence of non-relevant results further down the list were causing these limitations. While some of these literature directly evaluate the importance of a number of A&I databases, the others select A&I services to identify appropriate journals for literature studies across several subjects. In addition, we also found a few studies which were dedicated to perform citation analysis. However, despite the major objective of all these studies, one can easily assume that the influence and importance of corresponding A&I databases could lead to consider them in these studies.

The approximate percentage of representing each A&I database in all the 65 studies can be given as follows. Based on the percentages, it was possible to classify these databases into four separate categories and consequently to rank A&I databases.

- **15%≤Representation (Rank 1 databases)**
  - Web of Science - 19%
- **10%≤Representation<15% (Rank 2 databases)**
  - MEDLINE – 13%
  - Scopus – 10%
- **5%≤Representation<10% (Rank 3 databases)**
  - Embase, Google Scholar – 8%
  - PubMed – 6%
  - PsycINFO, CINAHL – 5%

- **Representation<5% (Rank 4 databases)**
  - Sociological Abstracts – 3%
  - Ulrich, Cochrane Library, PsycLIT - 2%
  - British Nursing Index, Social Works Abstracts, CAB Abstracts, Chemical Abstracts Service (CAS), CancerLIT, Applied Social Science Index and Abstracts (ASSIA), ERIC, Academic Search Elite, SCIRUS, GEOBASE, MLA International Bibliography – 1%
  - AGRIS, ProQuest, Gale, AMED, Zetoc, CNKI, AgeLine, ArticleFirst, CSA Illumina – 0.5%

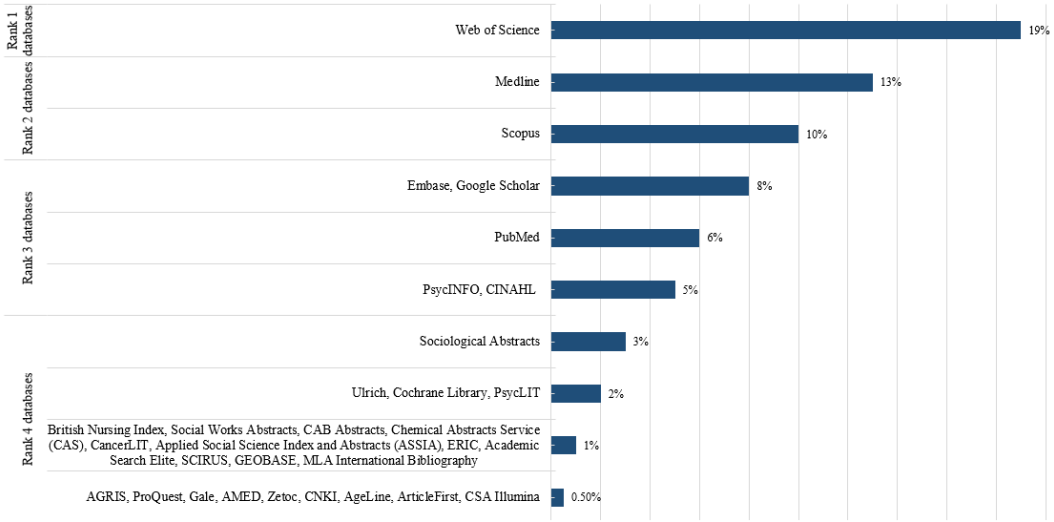


Figure 3.10: Percentages for popularity of A&I database

After ranking the identified A&I databases in the two subject domains, the criteria defined in table 3.4 was used to assign a numerical score for the journals which are indexed (or not) in numerous A&I databases.

Since all the considered journals are already indexed in the DOAJ, the following criterion does not consider the journal's indexation in DOAJ.

Criteria	Score
Indexed at least in one "Rank 1" database	4
Indexed at least in one "Rank 2" database	3
Indexed at least in one "Rank 3" database	2
Indexed at least in one "Rank 4" database	1
When none of the above cases are applied	0

Table 3.4: Scores for A&I databases

Missing data

Missing data is a common challenge faced by most research studies. Table 3.5 shows the percentages of the journals where it was difficult to find information even after sending a reminder email to the first email (see Appendix B) requesting the provision of metadata from editors/publishers. With the exception of three factors – number of full-text downloads, acceptance rates, and time from submission to first online appearance, information for all other factors was completed across all 947 journals.

Subject domain	No. of full-text downloads	Acceptance rates	Time from submission to first online appearance
Medicine	64%	8%	4%
Social sciences	69%	8%	5%

Table 3.5: Percentages of journals with missing data

The above figures report considerable data sparsity for a number of full-text downloads. The reasons for this were given by the journal editors in their responses:

1. Their publications are disseminated via mirror sites as well (e.g. PubMed, ProQuest, EBSCO, Gale Carnage, JSTOR). Therefore, the original editors and publishers are reluctant to produce download statistics reported only by their own journal website since the figures may not reflect the actual usage.

2. Some of the journals/publishers do not have technical facilities to collect download statistics.

3. Some journals do not provide downloadable formats like PDF. Numbers for read-only formats (e.g. HTML) are unreliable as they do not necessarily reflect

someone's interest with the content in any meaningful way. To illustrate, one can merely visit an HTML page to know whether the content is relevant. Thus, 'page views' is not a good measure.

4. There are journals which keep download counts on per IP basis (i.e. any number of downloads from the same IP address are considered as a single download).

5. Some experts about streaming on Internet argue that around half of the total flow on the web comes from robots. Therefore, some editors inform that all downloads are not made by humans.

6. Some journals provide download facilities only for the complete volume, but not for separate articles.

Moreover, the results of the first author survey (see section 4.1.1) that determined the importance given by the authors for 16 publishing factors, reported the least interest to the number of subscribers (i.e. number of full-text downloads). The corresponding mean importance values were 2.62 and 2.49 (according to 5-point Likert scale) in medicine and social sciences respectively. Therefore, considering these facts, we dropped the factor 'number of full-text downloads' from further study.

After dropping the factor 'number of full-text downloads', this study applied multiple imputation (Rubin, 1987) method to impute missing data in 'acceptance rate' and 'time from submission to online publication'. Among the other methods, multiple imputation receives higher attention due to its advantages like using for wide variety of scenarios (e.g. data missing completely, random missing, missing not at random) and ability of handling the uncertainty of imputations. Although a single imputation method like mean imputation would be easy to implement, it often causes biased estimates (Eekhout et al., 2012). The regression imputation limits the errors given by the mean imputation method, but it increases the risk of type I errors due to uncertainty about the imputed values (Enders, 2010). Methods like hot-deck imputation and last observation carried forward have their intrinsic faults (Wood et al.,

2004). SPSS software was used to apply multiple imputation with the following configurations:

1. Number of imputations – 10
2. Constraints –

   Time from submission to online appearance: minimum boundary was set to 0 (to avoid negative imputed values).

   Acceptance rate: minimum boundary was set to 0 (to avoid negative imputed values) and maximum boundary was set to 100 (to avoid exceeding 100% for acceptance rate).

3.5.2 A measure to determine similarity between author's criteria and journal's available criteria

Recommender systems use numerous measures to compare similarity between user's preference criteria regarding an item and true values available for the same criteria of the considered item. This process leads to decide the nearest neighbour available for the user and recommends an item accordingly. However, these measures and their aptness depend on factors like nature of the problem, dataset, filtering method, and so on. The current study used the Gower's measure (Gower, 1971) to determine the similarity between author's perception regarding journal selection criteria and the real values available for the same criteria of a journal. The Gower's measure is based on the Manhattan distance (Krause, 1973) and an adaptation of the Jaccard similarity coefficient (Jaccard, 1901). Its compatible characteristics including the handling of different variable types, measuring the similarity between two datasets, inclusion of separate weights for variables, simplicity and well recognition with collaborative filtering (Gräßer et al., 2016; Kagie et al., 2008; Pavoine et al., 2009) were the reasons for selecting Gower's similarity measure as an appropriate measure for the current problem. In addition, a number of studies have reported its flexibility

of usage in a wide variety of research areas. For example, recent works on ecological research (Chazdon et al., 2009; Olden et al., 2006; Pillar and Sosinski, 2003; Poff et al., 2006), medicine (Corrêa et al., 1999; Kosaki et al., 1996; Vitali et al., 1994), physics (Ogurtsov et al., 2002), and aquatic studies (Mansfield and Mcardle, 1998; Nascimento et al., 2000) can be given. Equation 3.14 represents the Gower's similarity measure.

$$S_{ij} = \frac{\sum_{k=1}^{n} w_{ijk} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}} \tag{3.14}$$

where S_{ij} is the overall similarity between i and j observations, while i and j are considered as the values an author expects for a factor that influences journal selection and the actual value attained by the published journal for the same factor. n is the number of variables and equals to the number of factors for the current study. w_{ijk} is the weight of the variable k when the similarity is measured between i and j observations, and finally s_{ijk} represents the similarity between i and j observations with respect to the variable k.

The similarity components of the equation 3.14 have been determined separately for different types of variables. The current problem has only binary type of variables in addition to ratio or interval type of variables.

Similarity for ratio or interval types of variables is determined by equation 3.15.

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{r_k}, \tag{3.15}$$

where, x_{ik} is the value of variable k for the observation i. Further, r_k is the range of the kth variable. This can be obtained by subtracting the maximum value of the variable k by the minimum value of the same variable.

Similarity for binary type of data can be determined by equation (3.16).

$$s_{ijk} = \begin{cases} 1, & \text{if } x_{ik} = x_{jk} = 1 \\ 0, & \text{if } x_{ik} = 0 \text{ or } x_{jk} = 0 \end{cases} \tag{3.16}$$

This implies the similarity between two observations becomes 1 if the variable value is present (we denote it as $x_{ik} = x_{jk} = 1$) for both observations. For example, the similarity between two observations becomes 1, if the considered journal is a peer-reviewed one while the author expects to submit a manuscript for a peer-reviewed journal. On the other hand, the similarity between two observations becomes 0, when one of the observations fails to present (i.e. $x_{ik} = 0$ or $x_{jk} = 0$).

Thereafter, equation 3.14 was implemented in the recommender system. The local Open Database Connectivity (ODBC) was established to receive (send) data from (to) the Microsoft Excel data file, in which the journal metadata was included. Also, the Java Database Connectivity (JDBC) driver was used to bridge the communication between the java content-based recommender application and the Excel database via ODBC (see Appendix H). This approach is relatively simple and could be faster than using a different database like SQL. Also, the database can be easily updated by a user who does not have knowledge of specific database programming.

Figure 3.11 illustrates the major components and operational procedure of the complete recommender system.

3.6 Second author survey: Collecting data to configure the recommender system

The objective of the second author survey was to collect authors' journal selection criteria for one of their articles published in 2018. It was aimed to input each criteria collected from authors to the recommender system with the corresponding article abstract to generate lists of appropriate journals. Therefore, a subsequent survey was planned to evaluate these journal lists with respect to the opinions of corresponding authors.

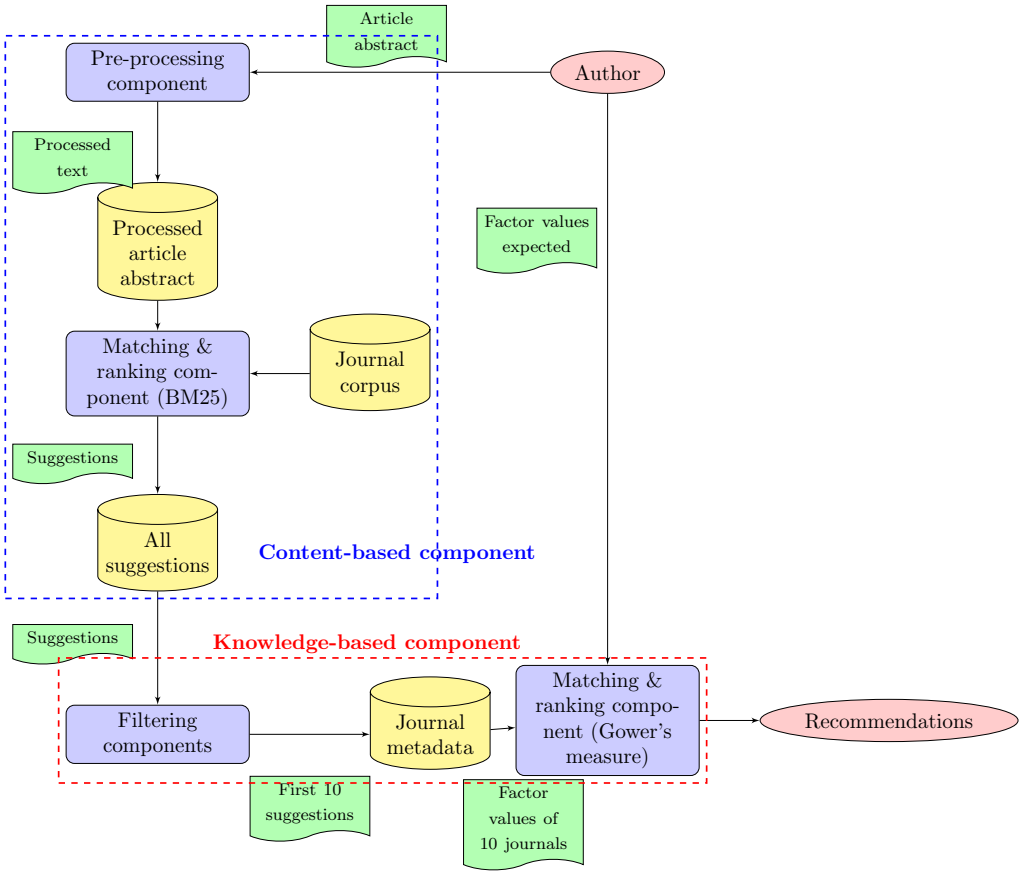


Figure 3.11: Architecture of the recommender system

3.6.1 Questionnaire and pre-test

As the survey instrument, we used a web-based questionnaire (see Appendix C.2) with a covering email invitation (see Appendix C.1) to the authors of journal articles. The questionnaire consisted of 15 key questions with two additional questions. The questions from one to four included only major questions which required simply yes/no answers from the respondents. They inquired whether the stated factors were considered while selecting an appropriate journal. The questions from five to fifteen allowed authors to select single or multiple options for sub-questions based on their yes/no choices to the given major questions.

One of the additional questions allowed respondents to input their further comments about the survey, while the other additional question requested to state their willingness to participate in the final survey of this study. Similar to the first author survey, the study used the LimeSurvey online survey tool to design, distribute, and administrate the web-based questionnaire.

Pre-test and amendments to the survey

This study conducted a pre-test before sending the questionnaire to the original participants of the survey. The main objectives of this pre-test were similar to the objectives of the pre-test of the first author survey described in section 3.3.2.

The questionnaire was sent to 9 participants and collected responses after 2 weeks from the date we sent the pre-test invitation email. Out of all participants, 8 completed the pre-test and submitted their comments. The implications of the comments are given below:

Some participants were not able to understand whether the survey aims to collect journal selection factors in common or with reference to a specific article which was published already. A few more journal selection factors were suggested for inclusion in addition to the current factors. For example, chief editor's, editorial board

members', and previous authors' titles with their affiliations. Availability of ISSN or E-ISSN and the amount of article processing charge were also suggested by the participants. Risk of receiving unintended answers for the questions 10 and 12 as it is difficult to find relevant information was highlighted. Less clarity of survey introduction including the survey objectives and target audience was a major failure of the pre-test. Some answer options were suggested to be more specific. For instance, peer-reviewed journals could be divided into categories like, open peer-reviewed, single blind, or double blind. Less clarity of answer options for the questions from 9 to 15 was highlighted by the respondents. Pre-test suggested further to modify the strategy of obtaining answers for questions from 9 to 15, as the existing method could be challenging for the survey participants.

Based on the comments received for the pre-test, a number of changes were made in the questionnaire. Introduction of the survey was elaborated further to make it easy to understand. Objectives were explained clearly. Moreover, the title of the article that each respondent has to consider while answering the questionnaire was stated in the email invitation. This strategy was expected to inform the respondents that the survey inquires journal selection factors regarding a particular article they have published, not about their journal selection criteria in general. The amount of author fee of each journal was included as a drop-down option in the questionnaire. We removed 'minimum', 'average', and 'maximum' answer options of questions from 9 to 15, as they may lead difficulties in understanding, given answer options and to avoid misleading answers. Instead of requesting exact values, they were changed to select answers from one or more categories (e.g. very recent, recent, middle-aged, etc. as answer options to journal's age). These categories were planned to map into appropriate values later. This amendment was done anticipating more responses as questions place less pressure on respondents, despite potential risk of receiving less precise answers. The second additional question, which inquires respondent's willingness to participate in the final survey of the study was introduced in the email invitation and introductory section of the survey too. This statement was expected to support respondents decision to participate in the current survey. Finally, a revision

of wording for the complete questionnaire was done to make the survey more clear.

3.6.2 Population, sample and data collection

The target of the second author survey was to draw a sample of authors from the same population of journals used by the study for constructing the document corpora described in section 3.4.2. The study followed purposive sampling method to select an appropriate author sample for the survey. Only the latest article of the latest issue of each journal was used to build the author sample. Special issues were skipped whenever they were encountered, as they could reflect only a specific aspect of the journal. However, special issues were replaced by the next most recent regular issue of the same journal. Only the corresponding author of each selected article was added to the sample. The first author was selected, whenever the corresponding author was not specified in the articles.

The following factors caused limitations in the author sample:

1. Inability of comparing classification results given by the articles from journals, which are not included in the journal corpora with articles from journals included in the corpora led to restriction of population of journals only from the already built corpora.

2. Journal metadata was collected only for the journals included in the system's journal corpora. This was the second reason to draw journals from the prebuilt journal corpora.

3. Authors from the most recent articles were considered in order to receive a higher response rate. For example, one can expect higher possibilities for valid author contact details, recalling journal selection criteria they used, and enthusiasm to respond, due to recent publications.

4. The corresponding author or the first author was contacted as he/she could

be the principal investigator who was mainly responsible for deciding the publication outlet.

5. Limited time availability to complete the study restricted the sample to a single article from each journal and only one author from each article.

6. We considered including all journals belonging to medicine and the social sciences to avoid possible similar answers that could be received, if the sample were to consist of authors from similar journals.

The author sample consisted of 530 medicine and 417 social sciences authors, corresponding to the number of medicine and social sciences journals included in the corpora of the recommender system. We collected the corresponding author's (or the first author's) full name, email address, and the article abstract for the two subject streams separately. The required information was collected manually from the official website of each journal or from the publisher's official website. Email addresses were located from the internet wherever the author's contact information was not mentioned in the article. Two lists of authors' names were arranged according to the alphabetical order of the author's last name. Any duplicate presence of the same author was replaced by another author of the same article. Whenever different authors with identical last names were found, they were sorted according to the name that appeared just before their last name. All sample authors were sent email invitations and requested to participate in the survey. The email message included a hyperlink to the survey database, allowing a direct connection to the questionnaire. Participants were allowed two weeks to answer the questionnaire before sending the reminder email. The survey was concluded after a week from sending the reminder email.

3.6.3 Mapping answer options

Mapping criteria

Answer options for questions from 7 to 12 of the second author survey were provided as nominal categories considering the comments received for the pre-test. Therefore, these categories were mapped into appropriate numerical categories for the convenience of data analysis. Moreover, this conversion was necessary to determine the distances from the factor values the authors stated as what they wanted to factor values of the journals they actually published in. Table 3.3 illustrates the procedure for determining numerical values for all factors of each journal belonging to medicine and social sciences subjects indexed in DOAJ. The least and highest values corresponding to each factor were considered as the factor's lower bound and upper bound respectively, and they were used to construct the numerical mapping categories for factors. For instance, the minimum SJR value of all social sciences journals in DOAJ (i.e. 0) was considered as the lower bound of the numerical scale of journal's prestige. Similarly, the upper bound of the same factor's scale was determined based on the maximum SJR value (i.e. 2.216) of the same collection of journals. Accordingly, the criteria in table 3.6 were defined for mapping each nominal category to the appropriate numerical category.

Numerical categories of IF, journal's prestige, publisher's prestige, and acceptance rate were obtained by dividing the corresponding ranges between lower bound and upper bound into five equal parts. Unavailability of a standard definition for nominal categories of these factors made it difficult to map nominal categories more specifically. Numerical categories of age and processing time were defined based on general appreciation of each factor. Two factors - number of issues and number of articles, which are numerically defined directly in the questionnaire were also followed using the same rule. For example, issues of a journal per year are usually furnished as bi-monthly, monthly, quarterly, bi-annually, and annually and this general notion was followed when defining numerical categories. In addition, numerical categories

Factor	Nominal category	Numerical category		Assigned value for category[a]	
		Medicine	Social sciences	Medicine	Social sciences
Age	Very recent	$1 < \text{Age} \leq 2$	$1 < \text{Age} \leq 2$	1.5	1.5
	Recent	$3 < \text{Age} \leq 5$	$3 < \text{Age} \leq 5$	4.0	4.0
	Middle-aged	$6 < \text{Age} \leq 20$	$6 < \text{Age} \leq 20$	13.0	13.0
	Old	$21 < \text{Age} \leq 50$	$21 < \text{Age} \leq 50$	35.5	35.5
	Very old	$51 < \text{Age} \leq 145$	$51 < \text{Age} \leq 114$	97.5	82.5
Processing time (PT)	Very short	$1 < \text{PT} \leq 3$	$1 < \text{PT} \leq 3$	2.0	2.0
	Short	$4 < \text{PT} \leq 12$	$4 < \text{PT} \leq 12$	8.0	8.0
	Average	$13 < \text{PT} \leq 24$	$13 < \text{PT} \leq 24$	18.5	18.5
	Long	$25 < \text{PT} \leq 48$	$25 < \text{PT} \leq 48$	36.5	36.5
	Very long	$49 < \text{PT} \leq 66$	$49 < \text{PT} \leq 53$	57.0	51.0
Impact factor (IF)	Very low	$0 < \text{IF} \leq 2.372$	$0 < \text{IF} \leq 1.259$	1.186	0.629
	Low	$2.372 < \text{IF} \leq 4.744$	$1.259 < \text{IF} \leq 2.518$	3.558	1.888
	Average	$4.744 < \text{IF} \leq 7.116$	$2.518 < \text{IF} \leq 3.777$	5.930	3.147
	High	$7.116 < \text{IF} \leq 9.488$	$3.777 < \text{IF} \leq 5.036$	8.302	4.406
	Very high	$9.488 < \text{IF} \leq 11.862$	$5.036 < \text{IF} \leq 6.296$	10.675	5.666
Journal's prestige (JP)	Very low	$0 < \text{JP} \leq 1.197$	$0 < \text{JP} \leq 0.443$	0.598	0.221
	Low	$1.197 < \text{JP} \leq 2.394$	$0.443 < \text{JP} \leq 0.886$	1.795	0.664
	Average	$2.394 < \text{JP} \leq 3.591$	$0.886 < \text{JP} \leq 1.329$	2.992	1.107
	High	$3.591 < \text{JP} \leq 4.788$	$1.329 < \text{JP} \leq 1.772$	4.189	1.550
	Very high	$4.788 < \text{JP} \leq 5.984$	$1.772 < \text{JP} \leq 2.216$	5.386	1.994
Publisher's prestige (PP)	Very low	$0 < \text{PP} \leq 1.197$	$0 < \text{PP} \leq 0.548$	0.598	0.274
	Low	$1.197 < \text{PP} \leq 2.394$	$0.548 < \text{PP} \leq 1.096$	1.795	0.822
	Average	$2.394 < \text{PP} \leq 3.591$	$1.096 < \text{PP} \leq 1.644$	2.992	1.370
	High	$3.591 < \text{PP} \leq 4.788$	$1.644 < \text{PP} \leq 2.192$	4.189	1.918
	Very high	$4.788 < \text{PP} \leq 5.984$	$2.192 < \text{PP} \leq 2.74$	5.386	2.466
Acceptance rate (AR)	Very low	$1 < \text{AR} \leq 20$	$1 < \text{AR} \leq 20$	10.5	10.5
	Low	$21 < \text{AR} \leq 40$	$21 < \text{AR} \leq 40$	30.5	30.5
	Average	$41 < \text{AR} \leq 60$	$41 < \text{AR} \leq 60$	50.5	50.5
	High	$61 < \text{AR} \leq 80$	$61 < \text{AR} \leq 80$	70.5	70.5
	Very high	$81 < \text{AR} \leq 100$	$81 < \text{AR} \leq 100$	90.5	90.5

[a] Middle value between minimum and maximum values of corresponding numerical category.

Table 3.6: Mapping from nominal to numerical categories

of international authorship were defined directly in the questionnaire, after dividing the range between the lower and upper bounds into five equal parts. Numerical categories provided for questions from 13 to 15 in the survey were also determined their 'Assigned value for category' based on the middle value between minimum and maximum values of corresponding numerical category.

Configuration

Equivalent article abstract of each respondent of the survey was identified using unique tokens passed with the hyperlink embedded in the email invitation. The content-based recommender component was first allowed to select appropriate journals based on the text similarities between input article abstract and the corpus articles. This step permitted the system to determine the appropriateness of journals based on similarity of subject area. The first 10 given results were further filtered by the knowledge-based recommender component based on author's publication needs. We used equation 3.14, which is implemented in the recommender system to compute similarity scores between the factor values the authors stated as what they wanted and the actual factor values for the 10 top-ranked journals retrieved by the content-based recommender component. Equation 3.14 was substituted by the values in column 'Assigned value for category' of table 3.6 (and middle values determined for questions 5 and from 13 to 15) to represent author's stated values for questions from 7 to 12, while actual factor values were determined from data collected in section 3.5.1. However, responses with multiple answers for single question (for questions from 7 to 15) were considered the middle value between lower bound of the lower numerical category and upper bound of the higher numerical category to substitute in equation 3.14. For instance, we calculated the middle value (7.116) between 4.744 (a lower bound) and 9.488 (an upper bound) of IF, when a medicine author selected both 'average' and 'high' as answer options for IF. Dichotomous (yes/no) answers received for survey questions from 1 to 4 were assigned 1 and 0 respectively, while answer for question 6 was substituted from table 3.4. Further, factor weights were

substituted from tables 4.4 and 4.5. Finally, the list of 10 journals was ranked based on the highest to least similarity scores.

The following working example illustrates the basic steps of this calculation when an author from social sciences considered only five factors to select an appropriate journal. Table 3.7 represents the first 10 journals retrieved by the content-based recommender component for author's input article abstract. Moreover, it shows the available factor values for each journal included in the list.

Imagine, for example, the author wants to publish the article in a journal, which has the following characteristics: peer-reviewed, belongs to a society, permanent article identifier, high IF, and with 6-8 issues per year. According to table 4.5, the five factors gain weights: 4.59, 3.13, 3.14, 3.77, and 2.86 respectively.

Journal	Peer-reviewed	Affiliation	Permanent identifier	IF	Issues/year
#1	Yes	No	Yes	1.515	6
#2	Yes	No	Yes	0.650	1
#3	Yes	Yes	Yes	1.029	4
#4	Yes	No	Yes	0	4
#5	Yes	No	Yes	0	3
#6	Yes	No	Yes	0	1
#7	No	No	Yes	0	4
#8	Yes	No	Yes	0	2
#9	Yes	Yes	No	0.270	2
#10	Yes	No	No	0	12

Table 3.7: Example - retrieved journals with values for 5 factors considered

Using these figures, similarity scores for 10 journals can be generated based on author's stated factor values (see table 3.8). Substituting values in equation 3.14 to calculate similarity score for journal #1 is illustrated below for further convenience of understanding.

$$S_{ij} = \frac{\sum_{k=1}^{5} w_{ijk} \times s_{ijk}}{\sum_{k=1}^{5} w_{ijk}}$$

$$S_{ij} = \frac{4.59 \times 1 + 3.13 \times 0 + 3.14 \times 1 + 3.77 \times [1 - \frac{|4.406-1.515|}{6.296}] + 2.86 \times [1 - \frac{|7-6|}{24}]}{4.59 + 3.13 + 3.14 + 3.77 + 2.86}$$

$S_{ij} = 0.7152500$

Where, i and j represent stated factor value by the author and actual factor value of the retrieved journal, while k= 1, 2, 3, 4, 5 denotes peer-review status, affiliation, permanent identifier, IF, and issues respectively.

Journal	Similarity score
#3	0.8639438
#1	0.7152500
#2	0.6515685
#4	0.6497552
#9	0.6448004
#5	0.6429418
#8	0.6361284
#6	0.6293150
#10	0.4565972
#7	0.3873195

Table 3.8: Example - ranked journals with similarity scores

3.7 Third author survey: Evaluating the recommender system

A succeeding survey to the second author survey was developed to compare the performance between only content-based recommender system and the knowledge-based recommender system together with the content-based recommender component[13]. This third author survey was directed to all respondents of the second author survey, those who gave consent to participate in the final step. The third author survey developed a table (see Appendix D.2) including a list of suggested journals generated by C&K recommender system according to its rank order. The survey table was developed using Google spreadsheet application[14] to facilitate participants to access the survey online. The email invitation (see Appendix D.1) to participants

[13]Thereafter, this finally merged recommender system (i.e. knowledge-based recommender component with content-based recommender component) is named as 'C&K recommender system' for the convenience of documentation. Here, C and K denote content-based and knowledge-based recommender components respectively.

[14]https://docs.google.com/spreadsheets/

included a hyperlink to connect with the online form. The main objective of the survey was to allow authors to rank the appropriateness (1-most appropriate to 10-least appropriate) of suggested journals for their articles based on their own opinion. In addition to a table with the answers participants were provided with for the second author survey, the main table of the survey included actual values that each journal attains for each journal characteristic that authors considered. This additional information was included with a view to keep survey participants more comfortable when deciding the ranks of journals listed in the third author survey.

This survey was pre-tested only with the principal advisor of the research. It was not necessary to test the survey with a large sample due to its direct and simple structure. Also, it is assumed that survey participants had a clear understanding about the survey, since research objectives were already explained in the second author survey. However, this survey was slightly adjusted according to the comments received. As a result, more important columns of the table were moved forward for the convenience of the survey participants, while adding a new column named 'subject of journal' to the table. This modification was expected to give a hint about subject scope of each journal listed in the survey. Furthermore, the specified ranking criteria was modified by adding 'N/A' as there could be not appropriate journals in the list for the considered article. Finally, the column headings of the tables were arranged horizontally to improve readability of the survey, in addition to re-wording the survey text appropriately.

There were 75 respondents from medicine and 113 respondents from the social sciences among those who participated in the second author survey and expected to participate in the third survey. All these authors were sent the email invitation to take part in the third author survey. They were allowed two weeks' time period to respond, before sending the reminder email. The third author survey was concluded a week after sending the reminder email.

This research used the Mann-Whitney U test, correlation analysis, PCA, chi-square test, and Fisher's exact test as the major data analyses methods. In addition to this,

the study also applied simple descriptive statistics to obtain and interpret the results. All analyses, except the polychoric correlation and PCA described in section 4.1, were performed using SPSS version 17. In order to determine the polychoric correlations and PCA results, this research used an exploratory factor analysis package called FACTOR[15], introduced by Lorenzo-Seva and Ferrando (2006).

[15] http://psico.fcep.urv.es/utilitats/factor/index.html

Chapter 4

Results

"We can do science, and with it we can improve our lives"

– Carl Sagan: *Cosmos* (1980), p.46

The current chapter of the dissertation elaborates the results obtained by the research following the methodology described in chapter 3. The chapter is organized into four major sections. Each section of the chapter describes the results including the details of data analysis. However, the discussion of the results and conclusions are included in chapter 5.

First, the importance of 16 journal factors from author's point of view is reported based on the data collected from the first author survey. The factors significantly differ in value with respect to the authors in the two subject domains and the correlations between the factors are also highlighted. Respondent's exposure to publishing field and current experience of using journal recommender systems are analyzed. PCA results are stated and the major categories into which the 16 journal factors partitioned are described in detail.

Second, the results on the performance of the five similarity algorithms are reported with respect to the test documents, sub-disciplines of journal corpora, average docu-

ment lengths of corpora documents, and the number of journals in each sub-discipline of two corpora. Based on these results, the most appropriate similarity algorithm for implementing the content-based recommender system is selected. Furthermore, these outcomes are used to interpret the behavior of the five algorithms based on the test documents and the nature of the corpora documents in the two distinct subject domains.

Third, the results obtained from the second author survey are reported and to what extent the authors considered the journal selection factors when they submitted one of their latest articles is analyzed. These basic results of the survey are used to configure the knowledge-based recommender system. The collected data from the survey is further used to determine the similarity between the overall factor values that authors stated in the survey as what they wanted and the actual factor values of the journals the article was published.

Fourth, the results obtained from the third author survey are described and the performance of the C&K recommender system against the author's opinion and the two gold standards is compared. Moreover, the performance of the C&K recommender system with the content-based recommender component alone is compared. Finally, some test results obtained for investigating the potential reasons for given performance differences between the C&K recommender system and the content-based recommender component alone are represented.

4.1 First author survey: Manuscript submission considerations

4.1.1 Major results

The survey conducted to collect information on the importance that authors assign for 16 journal factors received 129 and 106 fully completed replies from medicine

and social sciences authors respectively. Table 4.1 represents the basic statistics of the sample including bounced emails and response rates (Wijewickrema and Petras, 2017).

Sample	Medicine	Social sciences
Total sample	555	408
Emails not delivered	7	11
Effective sample	548	397
Completed survey	129	106
Response rate (approximately)	23.5%	26.7%

Table 4.1: Response rate of first author survey

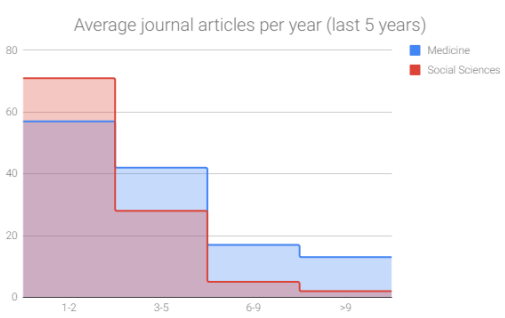
These response rates are higher than the 4% response rate received by Rowlands et al. (2004) and the 7.2% response rate obtained by Rowlands and Nicholas (2006) in their relatively large sample external web-based surveys conducted for collecting authors' perspectives regarding manuscript submission. This was a positive indication of the adequacy of response rate received by the current study. The percentages of authors' geographical distribution are given in table 4.2.

Geographical region	Medicine	Social sciences
Africa	7%	7%
Asia	25%	14%
Europe	33%	43%
Latin America and Caribbean	14%	23%
North America	16%	11%
Oceania	5%	2%

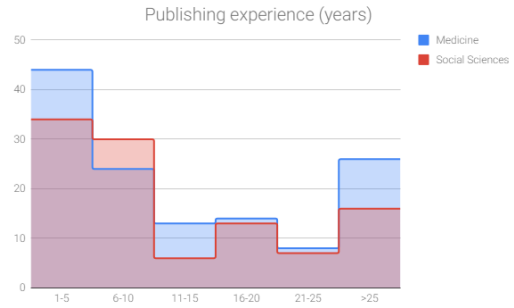
Table 4.2: Author percentages

Authors in the medicine subject domain represented 43 different countries while the number of representative countries for social sciences was 44. The highest response rate was reported by the authors of European countries. Authors from Oceania region showed the minimum interest of responding to the survey. Nevertheless, lack of information about geographical distribution of samples deter making a firm conclusion about the variation of authors' interests to respond to the survey based on their geographic region. Figure 4.1 visualizes the variation of three major characteristics

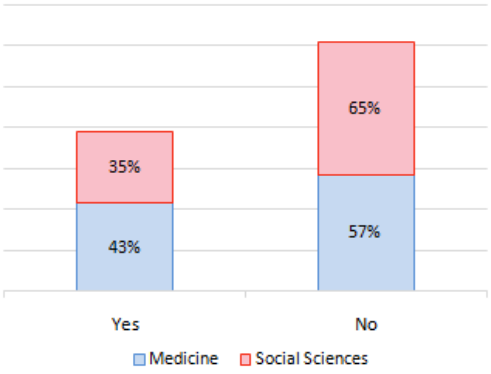
in the two subject domains. These three characteristics reveal author's publishing experience, recent contributions, and knowledge of publishing process as stated in section 3.3.1. Therefore, the output of these three characteristics can be considered as a collective indication of the author's exposure to the publishing process.



Average journal articles per year



Publishing experience



Editor/Editorial board member

Figure 4.1: Characteristics

Table 4.3 represents detailed percentages of these three characteristics of the respondents (Wijewickrema and Petras, 2017). The table indicates that about one third of the total respondents have 1 to 5 years of publishing experience, while the others have more than 5 years, showing that the sample authors have relatively higher experience in scholarly publishing. Authors in the social sciences publish less articles than authors in medicine. 67% of the social sciences authors published 1 to 2 articles per year during the past 5 years from 2017, while 44% of medicine authors produced only 1 to 2 articles. However, the percentage of medicine authors exceeds the social sciences authors when the number of articles during the last 5 years passes two articles per year. For example, 13% and 10% of medicine authors publish 6 to 9

and more than 9 articles respectively per year, yet the equivalent fractions are only 5% and 2% for the social sciences. This reflects a different research and publishing culture in the two subject domains (Wijewickrema and Petras, 2017). Perhaps, the higher dynamic nature of the science stream relative to social sciences may lead the medicine authors to publish more frequently than the other group.

A comparatively higher number of medicine authors (43%) work as editors or editorial board members compared to social sciences authors (35%). These figures indicate that the authors in both subject domains have sufficient experience of the publishing process.

Characteristic	Group	Frequency	
		Medicine	Social Sciences
Publishing experience (years)	1-5	44 (34%)	34 (32%)
	6-10	24 (19%)	30 (28%)
	11-15	13 (10%)	06 (06%)
	16-20	14 (11%)	13 (12%)
	21-25	08 (06%)	07 (07%)
	>25	26 (20%)	16 (15%)
Average journal articles per year (during last 5 years)	1-2	57 (44%)	71 (67%)
	3-5	42 (33%)	28 (26%)
	6-9	17 (13%)	05 (05%)
	>9	13 (10%)	02 (02%)
Editor or editorial board member	Yes	56 (43%)	37 (35%)
	No	73 (57%)	69 (65%)

Table 4.3: Percentages for respondents' characteristics

Then, the mean importance (weight) of each publishing aspect in deciding the publication outlet was calculated since the salient objective of the survey is devoted to determining the weights assigned by the authors for them. Tables 4.4 and 4.5 illustrate the mean importance separately, for the domains of medicine and social sciences respectively. In addition, they show the corresponding number of responses received by each Likert scale level for all the 16 factors. The Likert scale is defined from rate 1=not important at all to rate 5=very important.

The Mann-Whitney U test was used to check for significant differences in the importance of the 16 factors between the medicine and social sciences authors. The $p-$value

Factor	No. of responses for					Weight
	rate 1	rate 2	rate 3	rate 4	rate 5	
Peer-reviewed	01	02	08	20	98	4.64
Impact Factor (IF)	00	01	19	47	62	4.32
Journal's Prestige	01	04	26	52	46	4.07
Publisher's prestige	08	18	35	44	24	3.45
Representing society	16	21	35	40	17	3.16
No. of subscribers per year	26	36	39	17	11	2.62
Abstracting & indexing	00	08	12	32	77	4.38
Author contributions	21	19	41	31	17	3.03
Persistent identifier	11	21	31	33	33	3.43
Age of journal	19	37	45	20	08	2.70
No. of issues per year	23	40	30	25	11	2.70
Processing time	03	13	44	40	29	3.61
Acceptance rate	02	17	51	36	23	3.47
Online submission/tracking	07	08	28	54	32	3.74
No. of papers per year	15	33	38	28	15	2.96
No author charges	06	12	28	26	57	3.90

Table 4.4: Responses received for factors and their weights in medicine

Factor	No. of responses for					Weight
	rate 1	rate 2	rate 3	rate 4	rate 5	
Peer-reviewed	01	01	11	14	79	4.59
Impact Factor (IF)	04	11	27	27	37	3.77
Journal's Prestige	00	13	19	35	39	3.94
Publisher's prestige	10	26	26	31	13	3.10
Representing society	13	22	28	24	19	3.13
No. of subscribers per year	30	26	24	20	06	2.49
Abstracting & Indexing	10	09	18	20	49	3.84
Author contributions	18	15	28	22	23	3.16
Persistent identifier	20	13	27	24	22	3.14
Age of journal	20	32	30	19	05	2.59
No. of issues per year	22	20	30	19	15	2.86
Processing time	02	14	29	29	32	3.71
Acceptance rate	03	19	38	27	19	3.38
Online submission/tracking	08	21	28	25	24	3.34
No. of papers per year	24	21	26	22	13	2.80
No author charges	06	09	19	25	47	3.92

Table 4.5: Responses received for factors and their weights in social sciences

results (see table 4.6) (Wijewickrema and Petras, 2017) show that the IF, publisher's prestige, abstracting & indexing, and online submission with tracking facility are the factors, which significantly differed (at the 5% level) between the authors in the medicine domain from the authors in social sciences. Here, the U value represents the number of times observations in one subject domain precede observations in the other sample in ranking. The Mann-Whitney test reports information like Mann-

Whitney U value in addition to the $p-$value since additional information could be useful for further studies (Hart, 2001). The $p-$value indicates whether medicine and social sciences are selected from populations having the same distribution. In other words, the $p-$value can be used to decide the existence of significant differences between the two subject domains. Usually, the cases with $p-$value<0.05 are referred to as significantly different between the two samples (Tallarida and Murray, 1987).

Factor	Mann-Whitney U	$p-$value
Peer-reviewed	6693.5	0.714
IF	5089.5	0.000345
Journal's prestige	6510	0.505
Publisher's prestige	5707	0.025
Representing society	6732	0.835
No. of subscribers per year	6421	0.409
Abstracting & Indexing	5442.5	0.003
Author contributions	6439	0.431
Persistent identifier	6055	0.122
Age of journal	6479	0.475
No. of issues per year	6364	0.350
Processing time	6452	0.440
Acceptance rate	6502	0.500
Online submission/tracking	5533	0.009
No. of papers per year	6363	0.349
No author charges	6741.5	0.846

Table 4.6: U test for factors' importance

We applied polychoric correlations (Pearson and Pearson, 1922) to determine the association between the factors, because variables measured on a Likert scale do not follow the main assumptions of the Pearson or Spearman correlations since they usually do not follow linear or monotonic relationships. Table 4.7 (Wijewickrema and Petras, 2017) depicts that there exist fairly strong positive correlations with values greater than 0.5 (Priebe et al., 1996) between two and three pairs of factors in the medicine and social sciences domains respectively. To illustrate, for example, if the 'number of journal issues per year' receives a higher importance, it influences to increase the importance of the factor 'number of papers per year' and vice versa. Moreover, this condition is valid for both subject domains independently.

Pair of factors	Correlation	
	Medicine	Social sciences
Author contributions & Persistent identifier	0.504	0.606
No. of issues per year & No. of papers per year	0.536	0.594
No. of papers per year & Online submission/tracking	<0.5	0.522

Table 4.7: Important correlations

The study discovered the authors' awareness of existence of journal recommender systems and to what extent the existing systems are used by them. Respondents' attitudes about usefulness of journal recommender systems were also collected to examine their impact on researchers. Figure 4.2 offers a quick visualization of the collected statistics, while table 4.8 gives respondents' detailed statistics of their experience of using journal recommender systems along with how they view their usefulness. Approximately 35% of the total respondents of each subject domain were aware of the existence of journal recommender systems. However, out of the total, 10% and 11% authors in the medicine and social sciences domains respectively, do not use journal recommender systems at all for selecting appropriate publication outlets. The majority of the respondents, only who were aware of the journal recommender systems, use them either often (4% in each subject domain) or sometimes (21% in medicine and 20% in social sciences) to select appropriate journals to submit articles. A significant percentage (67%) of respondents in medicine and more than half (55%) respondents in social sciences are of the view that journal recommender systems could assist authors to select appropriate journals for articles. Only 8% in medicine and 11% in social sciences held the view that the journal recommender systems could not help authors to find fitting journals – while 24% of medicine authors and 34% of authors in social sciences have no clear idea about the usefulness of these systems.

Subject domain	Experience of authors who aware			Usefulness		
	Often	Sometimes	Not at all	Helpful	Neutral	Not helpful
Medicine	5 (4%)	27 (21%)	13 (10%)	87 (67%)	31 (24%)	11 (08%)
Social Sciences	4 (4%)	21 (20%)	12 (11%)	58 (55%)	36 (34%)	12 (11%)

Table 4.8: Author percentages for their experience and usefulness of journal recommender systems

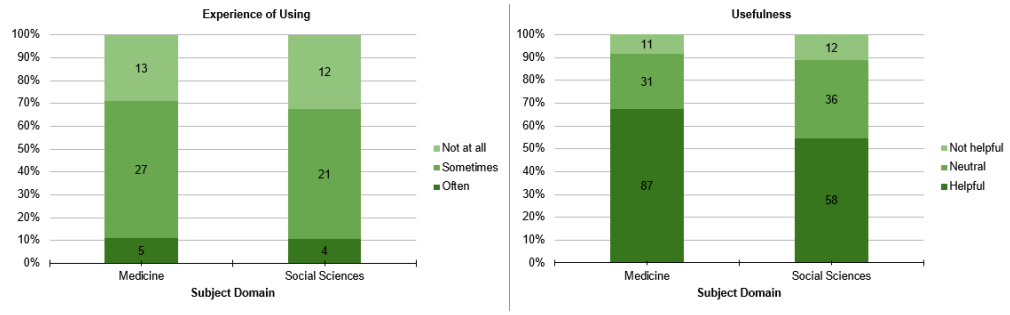


Figure 4.2: Experience of authors who aware of recommender systems and usefulness

4.1.2 Further results of first author survey

This study conducted PCA to identify the major groups that summarize the list of 16 journal factors (Wijewickrema and Petras, 2017). PCA was applied separately to the two subject domains. First, the 16 factors were projected based on two to five components separately. However, the results gave a meaningful classification for only three components. These three components are described in section 5.1. The PCA followed Kaiser's criterion (Kaiser, 1974; Kaiser and Rice, 1974) and Bartlett's test (Bartlett, 1950) before performing the orthogonal rotation. Usually, the Kaiser-Meyer-Olkin (KMO) value determines the suitability of data for PCA. KMO returns values between 0 and 1, and the criteria in table 4.9 is used to interpret the sampling adequacy for applying PCA (Kaiser, 1974). The sample adequacy and appropriate-

Kaiser-Meyer-Olkin (KMO) value	Appropriateness for PCA
0.00 to 0.49	Unacceptable
0.50 to 0.59	Miserable
0.60 to 0.69	Mediocre
0.70 to 0.79	Middling
0.80 to 0.89	Meritorious
0.90 to 1.00	Marvelous

Table 4.9: Interpretation of KMO values for PCA

ness of applying PCA for the current problem were confirmed by the test results as KMO value is greater than 0.5 and the $p-$value of Bartlett's test is less than 0.05 (Williams et al., 2010). Test results obtained are shown in table 4.10 (Wijew-

ickrema and Petras, 2017).

Test	Medicine	Social sciences
Kaiser-Meyer-Olkin (KMO)	0.83902	0.85711
Bartlett	825.5 ($p = 0.000010$)	791.4 ($p = 0.000010$)

Table 4.10: Test results

Varimax orthogonal rotation (Kaiser, 1958) allowed to reduce the number of loadings on each component, while identifying the most important loadings. As Field (2013) explains, usually, the loading value 0.3 is acknowledged as the minimum boundary to consider the most important loadings. Therefore, all loadings smaller than 0.3 were suppressed in table 4.11 (Wijewickrema and Petras, 2017) and it led to interpret results appropriately.

Factor	Medicine			Social sciences		
	Component					
	1	2	3	1	2	3
Peer-reviewed	0.731			0.731		
IF	0.545			0.629		
Journal's prestige	0.672			0.736		0.356
Publisher's prestige	0.462		0.518	0.515		0.509
Represent institute/society			0.572			0.553
Subscribers per year			0.716			0.668
A&I	0.585	0.353		0.387	0.387	0.327
Author contributions			0.639			0.704
Persistent identifier	0.458		0.448			0.690
Age of journal			0.680			0.645
Issues per year		0.390	0.637		0.637	0.477
Time: submission-publish		0.747			0.760	
Acceptance rate		0.613	0.322	0.304	0.630	
Online with tracking		0.645			0.604	0.368
Articles per year		0.538	0.622		0.624	0.492
No author charges		0.642		0.360	0.391	

Table 4.11: Loadings on three components

According to table 4.11, it is clear that each factor loads at least on one common component in both subject domains. This leads to the conclusion that the authors of both medicine and the social sciences recognize these 16 journal factors similarly.

Thus, table 4.12 (Wijewickrema and Petras, 2017) arranges the factors into three components considering only those factors, which were categorized in these components in both domains.

Component 1	Component 2	Component 3
Peer-reviewed	Processing time	Representing society
IF	Acceptance rate	Subscribers per year
Journal's prestige	Online with tracking	Author contributions
	No author charges	Persistent identifier
		Age of journal
A&I	A&I	
Publisher's prestige		Publisher's prestige
	Issues per year	Issues per year
	Articles per year	Articles per year

Table 4.12: Common loadings

At the end of the first author survey, the average importance authors allocate to each of the three components were determined. The mean importance of each factor was first averaged over subject domain and then averaged by component. Accordingly, component 1 gained the highest importance (average 4.01 out of 5 point scale), while average importance values 3.47 and 2.95 were achieved by component 2 and component 3 respectively. A meaningful interpretation for these three components is given in section 5.1.

4.2 Content-based recommender system

Each test document (i.e. abstract with title and keywords) from the two separate samples with 179 medicine and 164 social sciences documents described in section 3.4.4 was fed as input to the system. The system considered all key terms of the test document to compare with the corpus documents. A list of journal titles that mostly related to the input test document was given as the output of the recommender system. Figure 4.4 shows the list of journal titles suggested by the system for an input journal article abstract (see figure 4.3), which belongs to the sub-discipline 'Eco-

nomics', while using the cosine algorithm as the similarity measure. It depicts that the system arranges the 10 most appropriate journals based on the similarity values they scored. The journals at rank positions 1-5, 8, and 10 of the output (see figure

Macroeconomic determinants of the labour share of income: Evidence from OECD economies

The study investigates the relationships between the labour share of income and several macroeconomic variables - the GDP growth, inflation, unemployment, as well as GDP gap and capacity utilization - in industrialised economies between 1960 and the 2010s. Three complementary hypotheses that relate macroeconomic determinants to the labour share dynamics are considered: 'overhead labour' hypothesis, 'realization theory/wage lag' hypothesis and the 'rising strength of labour' hypothesis. The study employs a sequential procedure: testing for the stationarity properties of the variables, using bounds test to identify the presence of cointegrating relationships, and estimating long-run relationships using ARDL or OLS methods. The results show that all three hypotheses are supported only in a limited number of economies, whilst in the majority of cases only certain relationships are prominent. On the whole, the GDP growth rate, the unemployment rate, and to a smaller extent capacity are found to be the principal determinants of the labour share, while change in the level of prices is of subsidiary importance.

Keywords: labour share; time series; macroeconomic determinants

Figure 4.3: Input abstract from "Economics"

```
List of Appropriate Journals:

[1] Revista Economica
[2] Global Economic Observer
[3] Socioeconomica
[4] The Romanian Economic Journal
[5] EuroEconomica
[6] Journal of Innovations and Sustainability
[7] Technology Audit and Production Reserves
[8] Proceedings of Rijeka Faculty of Economics
[9] Africa Spectrum
[10]Review of Economics & Finance
```

Figure 4.4: Output of the content-based system

4.4) belong to the sub-disciplines 'Economics', while the journals at rank positions 6, 7, and 9 belong to the sub-disciplines 'Sustainable development', 'Business', and 'Social sciences' respectively.

4.2.1 Performance of algorithms against test documents

Table 4.14 summarizes the average NDCG values calculated for all test documents for each similarity measure and classification algorithm in two subject domains. According to the table, BM25 gives the highest average NDCG value for test documents in both subject domains while unigram language measure reports the lowest. Moreover, the corresponding average values given by each measure in two subject domains are approximately similar to each other. In general, NDCG value ranges from zero to one.

Equation 3.11 was used to calculate the NDCG for the test documents.
For example, assume that the results given in table 4.13 were retrieved by an input test document. The relevance of each retrieved journal to the test document is determined by comparing the sub-discipline category of the test document and the retrieved journal according to the method described in section 3.4.4. The information

Position of the retrieved journal (i)	Relevance (rel_i)
1	4
2	4
3	3
4	4
5	4
6	3
7	1
8	2
9	1
10	0

Table 4.13: Example - Computing NDCG

in table 4.13 can be used to calculate the DCG component defined in equation 3.12. We use $p = 10$, since the first 10 results are considered.

$$DCG_{10} = rel_1 + \sum_{i=2}^{10} \frac{rel_i}{\log_2 (i+1)}$$

$$DCG_{10} = 4 + \frac{4}{\log_2 3} + \frac{3}{\log_2 4} + \frac{4}{\log_2 5} + \frac{4}{\log_2 6} + \frac{3}{\log_2 7} + \frac{1}{\log_2 8} + \frac{2}{\log_2 9} + \frac{1}{\log_2 10} + \frac{0}{\log_2 11}$$

$DCG_{10} = 13.6$

Equation 3.13 can be used to calculate the IDCG component, since we already know the number of journals belonging to each sub-discipline category of the corpus. Assuming there are five journals appropriate for the test document with the relevance 4, two journals with the relevance 3, and at least three journals with the relevance 2, IDCG can be determined as follows:

$$IDCG_{10} = 4+\frac{4}{\log_2 3}+\frac{4}{\log_2 4}+\frac{4}{\log_2 5}+\frac{4}{\log_2 6}+\frac{3}{\log_2 7}+\frac{3}{\log_2 8}+\frac{2}{\log_2 9}+\frac{2}{\log_2 10}+\frac{2}{\log_2 11}$$

$IDCG_{10} = 15.6$

Therefore, using equation 3.11, NDCG value is obtained as 0.871.

Subject domain	Unigram Language	BM25	Cosine	SVM	MNB
Medicine	0.121	0.615	0.469	0.218	0.354
Social Sciences	0.098	0.626	0.436	0.139	0.313

Table 4.14: Average NDCG in two subject domains (179 medicine and 164 social sciences test cases)

4.2.2 Performance of algorithms against sub-discipline

Figures 4.5 and 4.6 give the variation of average NDCG values based on the sub-disciplines in the two subject corpora. As stated in section 3.4.4, the test documents in medicine belong to 38 sub-disciplines while the test documents in social sciences belong to 31 sub-disciplines (based on the LCC classification categories). However, some of the test documents belong to more than one sub-discipline and this property is common to both subject domains. BM25 scores better on average than cosine similarity and MNB for 96.8% of the sub-disciplines included in the social sciences corpus. Cosine similarity performs better than BM25 for only 'Technology' sub-discipline (Wijewickrema et al., 2019). MNB outperforms BM25 for only Economics sub-discipline. Cosine similarity measure performs better than MNB for 83.9% of sub-disciplines. The unigram language measure and SVM give relatively low NDCG

values for most sub-disciplines in contrast to the other three algorithms. Moreover, one can notice that neither unigram language measure nor SVM outperform BM25 or cosine measure for at least a single sub-disciplinary category in the social sciences. The situation is somewhat different in the medicine subject domain. There are three sub-disciplines – 'Anesthesiology', 'Otorhinolaryngology', and 'Tropical medicine', in which cosine similarity outperforms BM25. Cosine similarity does not exceed the performance of the BM25 for any other sub-discipline. The BM25 works better than cosine similarity for approximately 92.1% of the total number of sub-disciplines in the medicine corpus (Wijewickrema et al., 2019). Although, MNB does not perform better than BM25 for any of the sub-disciplines under the medicine, it performs better than the cosine measure for approximately 13% sub-disciplines. The performance of SVM is slightly improved in the medicine domain with respect to the cosine measure, because 7.9% of sub-disciplines report better NDCG for SVM than the cosine measure. However, the behavior of unigram language measure compared to both BM25 and cosine measure is approximately similar in both corpora.

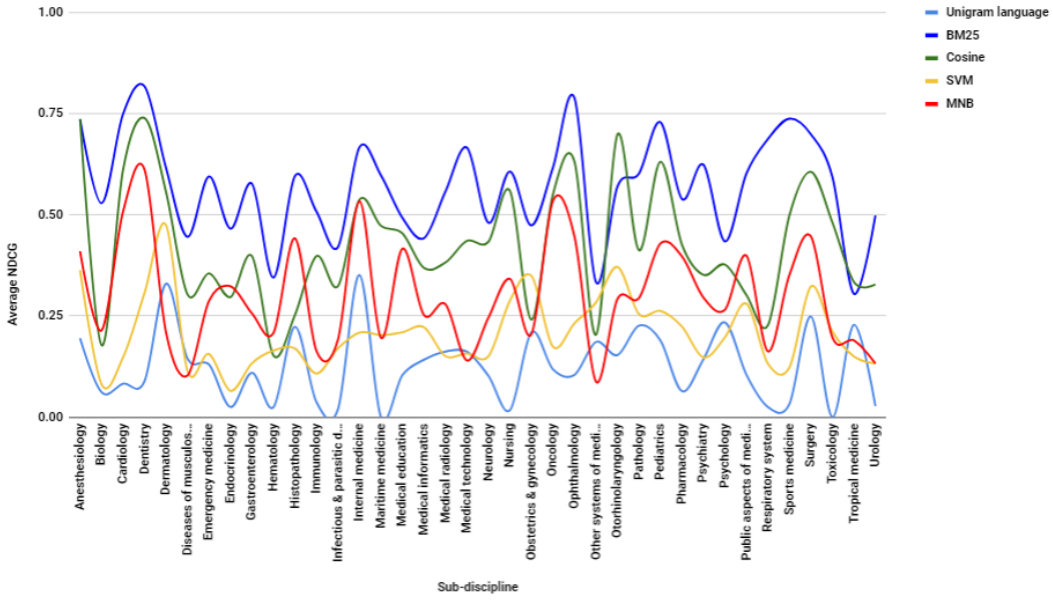


Figure 4.5: Average NDCG values against sub-disciplines in medicine

Mann-Whitney U test was applied to check for statistically significant differences in performance of the five algorithms based on the average NDCG values scored in each sub-discipline (see tables 4.15 and 4.16). The results showed that there

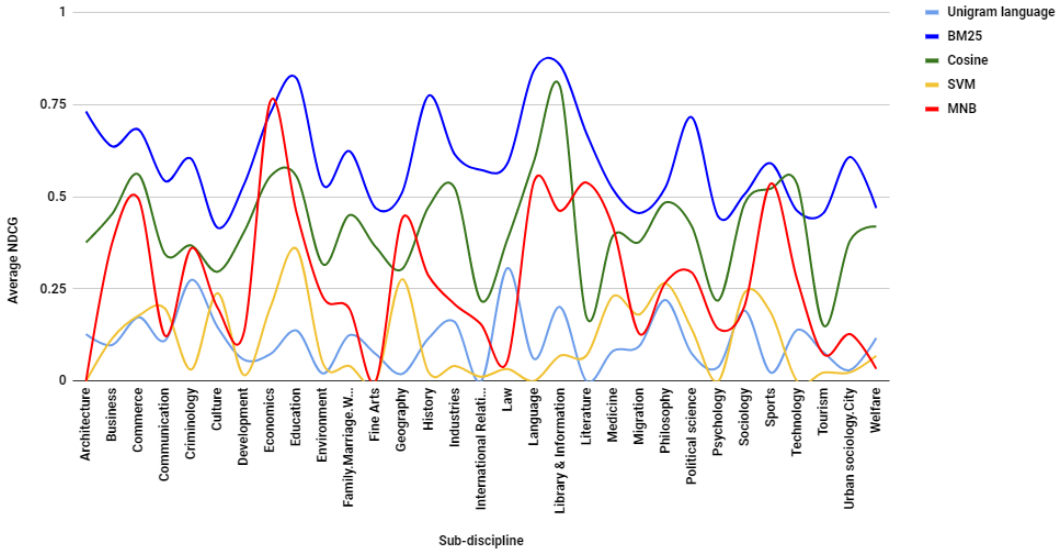


Figure 4.6: Average NDCG values against sub-disciplines in social sciences

exists a significant difference (at the 5% level) of sub-discipline wise averaged NDCG scores for each pair of algorithms in both subject domains, except between unigram language model and SVM in the social sciences.

Algorithm	BM25		Cosine		SVM		MNB	
	M-W*	p-value	M-W	p-value	M-W	p-value	M-W	p-value
Unigram	4	0.000	54	0.000	370	0.000	191	0.000
BM25			342	0.000	20	0.000	106	0.000
Cosine					143	0.000	386	0.000
SVM							413	0.001

* Mann-Whitney U statistic

Table 4.15: p-values obtained for pairs of algorithms in medicine

Algorithm	BM25		Cosine		SVM		MNB	
	M-W	p-value	M-W	p-value	M-W	p-value	M-W	p-value
Unigram	0	0.000	20	0.000	434	0.512	210	0.000
BM25			146	0.000	0	0.000	81	0.000
Cosine					38	0.000	253	0.001
SVM							214	0.000

Table 4.16: p-values obtained for pairs of algorithms in social sciences

Moreover, this study applied Spearman correlation test to find the existence of potential correlations between the performance of the algorithms in sub-disciplines. The obtained results (tables 4.17 and 4.18) explore the existence of moderate ($> |0.50|$ and $< |0.70|$) (Mukaka, 2012), but statistically significant correlations of performance

between each pair of BM25, cosine, and MNB algorithms in the sub-disciplines of both subject domains. Potential reasons for these behaviors are given in section 5.2.

Algorithm	Unigram	BM25	Cosine	SVM	MNB
Unigram	1.000	0.067	0.119	0.382*	0.184
BM25		1.000	0.672**	0.230	0.615**
Cosine			1.000	0.478**	0.575**
SVM				1.000	0.324*
MNB					1.000

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Table 4.17: Correlation between algorithms in medicine sub-disciplines

Algorithm	Unigram	BM25	Cosine	SVM	MNB
Unigram	1.000	0.089	0.436*	0.184	−0.086
BM25		1.000	0.544**	−0.009	0.514**
Cosine			1.000	0.188	0.510**
SVM				1.000	0.412*
MNB					1.000

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Table 4.18: Correlation between algorithms in social sciences sub-disciplines

4.2.3 Influence of average document lengths of training corpus

Figure 4.7 depicts the variation of average document lengths of corpus journal articles based on the sub-disciplines in each subject domain (Wijewickrema et al., 2019). The study defined the document length of a corpus document as follows:

The total number of words in the title, keywords, abstract, body, and references of an article in a corpus.

Calculating the document length was based on the full-text since the corpora included full-text articles. We can clearly notice that the average document lengths of the social sciences articles are higher than the average document lengths of the medicine articles.

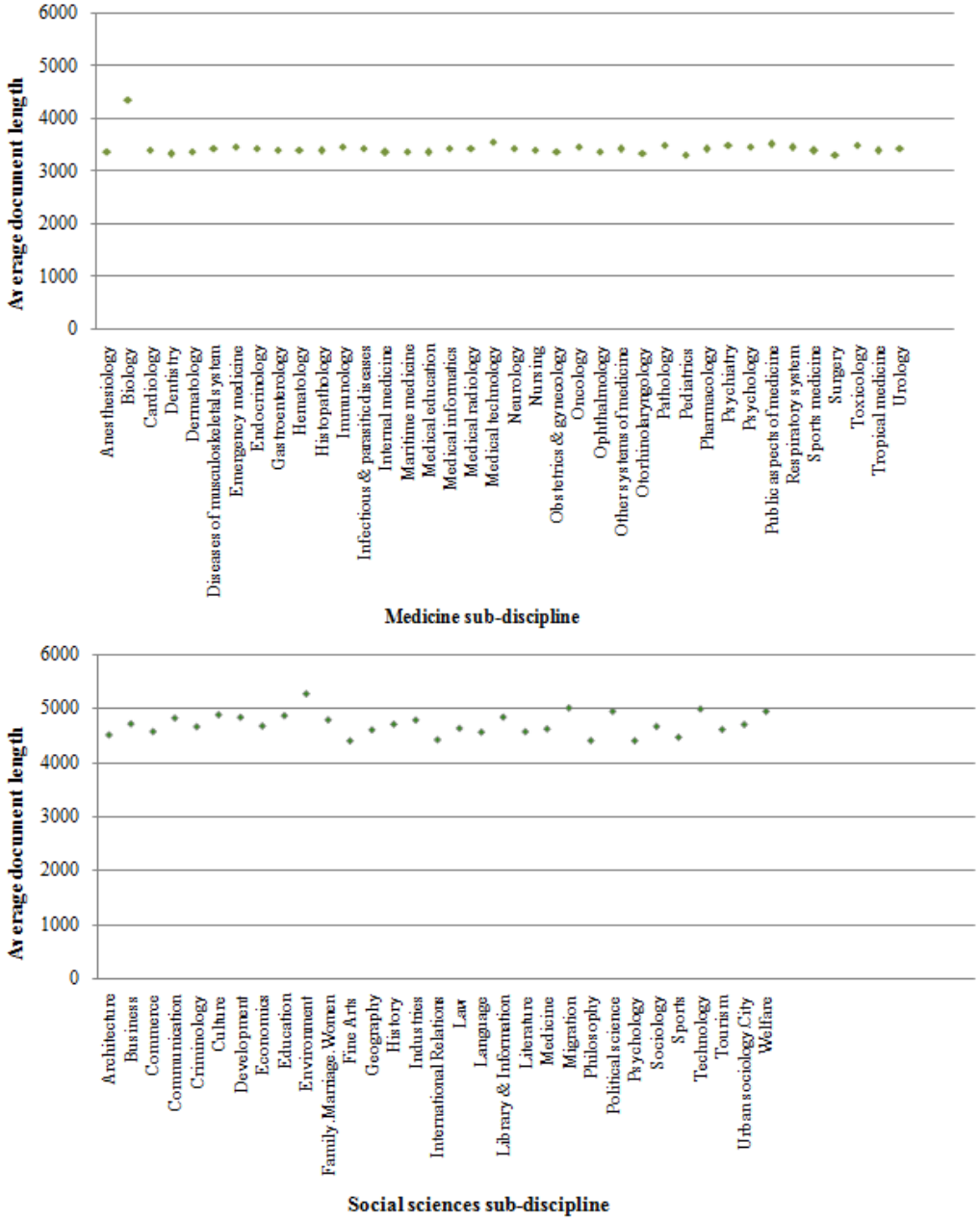


Figure 4.7: Average article lengths of journals in two training corpora

The Spearman correlation test was applied to determine the relationship between the average lengths of journal articles belonging to sub-disciplines in both subject domains and the performance of the five algorithms. Table 4.19 shows the correlation results obtained for each algorithm in the medicine and social sciences subject domains.

Table 4.19 reveals that the only moderate, but statistically significant correlations

Subject domain	Unigram	BM25	Cosine	SVM	MNB
Medicine	−0.244	−0.274	−0.531*	−0.504*	−0.417*
Social sciences	0.217	−0.024	0.181	0.211	−0.050

* Correlation is significant at the 0.01 level (2-tailed).

Table 4.19: Correlation for average document lengths and algorithms

are observed between document lengths and the performance of two algorithms in the medicine domain. These correlations are given by the cosine and SVM algorithms. All the other cases have low or negligible correlations ($< |0.50|$) between the average document lengths of sub-disciplines and the text similarity measures or supervised classification algorithms. In addition, table 4.19 shows that document lengths in medicine have slightly higher values for correlations than in the social sciences domain.

4.2.4 Influence of number of journals belonging to sub-disciplines of training corpus

Figures 4.8 and 4.9 show the number of corpus journals belonging to each sub-discipline of the two training corpora.

The Spearman correlation test was applied to reveal the relationship between the number of corpus journals per category and the performance of five algorithms. Results are given in table 4.20.

Subject domain	Unigram	BM25	Cosine	SVM	MNB
Medicine	−0.111	0.300	0.199	−0.121	0.311
Social Sciences	0.184	0.442*	0.546**	0.334	0.601**

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Table 4.20: Correlation for number of corpora journals and algorithms

According to the results obtained in table 4.20, there is a negligible correlation between the number of sub-disciplinary journals and the performance of the unigram

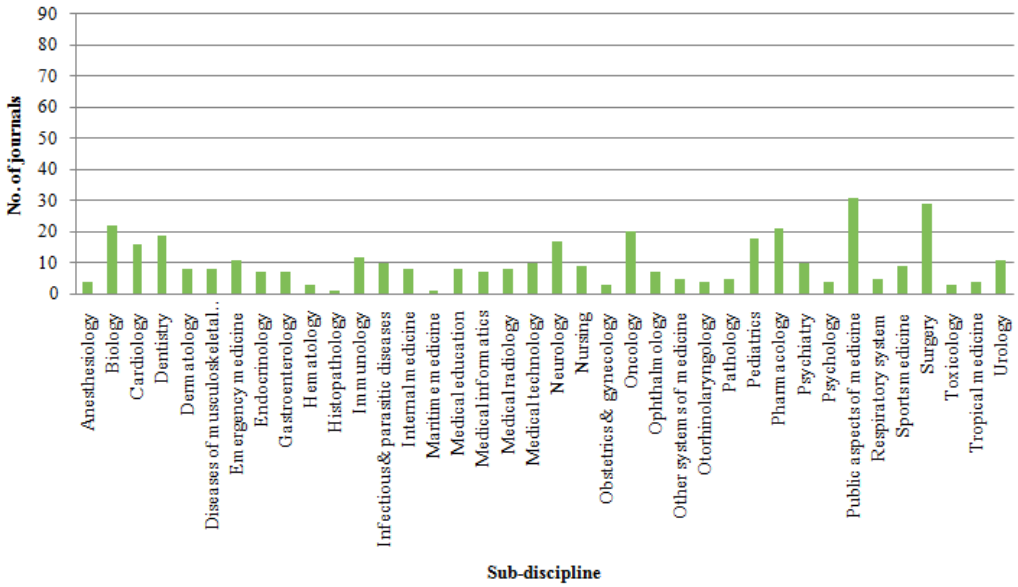


Figure 4.8: Number of journals belonging to sub-disciplines of medicine training corpus

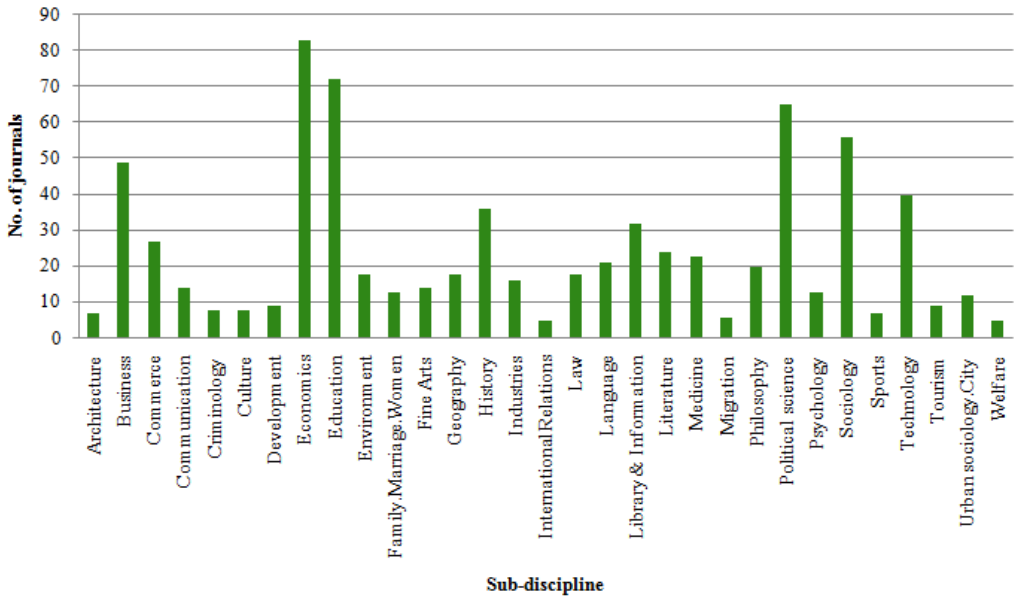


Figure 4.9: Number of journals belonging to sub-disciplines of social sciences training corpus

language measure in both subject domains. In the social sciences subject domain, BM25 reveals low ($>|0.30|$ and $<|0.50|$), statistically significant, and positive correlation while the correlation is low and positive in the medicine domain. The cosine similarity gives a moderate, statistically significant, and positive correlation in the

social sciences subject domain, but a negligible correlation in the medicine subject domain with the number of journals (Wijewickrema et al., 2019). The behavior of the SVM is closer to unigram language measure, but shows a low correlation with the number of sub-disciplinary journals in the social sciences domain. MNB behaves similar to cosine measure in the social sciences domain, but correlates similar to BM25 in the medicine domain.

4.3 Second author survey: Collecting data to configure the recommender system

4.3.1 Descriptive statistics

The second author survey was received responses from 75 medicine authors and 113 social sciences authors. All respondents had agreed to participate in the third author survey. Prior awareness of a succeeding survey possibly helped sample authors' decision of participating in the second author survey. Table 4.21 shows the summary of sample and respondents of the survey.

Sample	Medicine	Social Sciences
Total sample	530	417
Emails not delivered	12	15
Effective sample	518	402
Completed survey	75	113
Response rate (approximately)	14.5%	28.1%

Table 4.21: Response rate of second author survey

The following characteristics of the survey participants in both subject domains were observed based on the answers provided. Table 4.22 gives the approximate percentages of authors who considered each of the 15 journal factors as appropriate to consider for submitting their articles.

Table 4.23 gives the author fee categories which were acceptable to authors – those

Factor	% of authors who considered factor	
	Medicine	Social Sciences
Peer-reviewed	96.0	94.7
IF	80.0	53.1
Journal's prestige	89.3	78.8
Publisher's prestige	52.0	58.4
Presence of an affiliation	53.3	63.7
Abstracting & Indexing	88.0	73.5
International authorship	52.0	55.8
Permanent article identifier	61.3	61.9
Age of journal	24.0	23.0
Number of issues per year	29.3	28.3
Processing time	66.7	65.5
Acceptance rate	41.3	44.2
Online submission with tracking	85.3	66.4
Number of articles per issue	18.7	24.8
No author charges	60.0	76.1

Table 4.22: Author percentages for considered factors

who did not necessarily consider journals with only free of author charges. The approximate percentages of authors who belong to each category are given. Moreover, it depicts author percentages for indexing and abstracting services suggested by the survey. However, this table does not include A&I options for percentage values less than 5% in both subject domains.

Table 4.24 shows the variation of author percentages for answer options belonging to journal factors from 7 to 15 in the survey. It is important to note that the total percentage of each factor exceeds 100 since respondents were allowed to select more than one answer option for the same factor. Further interpretation of the results obtained is included in section 5.3.

4.3.2 Further results

Studying the nature of 15 journal factors in two subject domains is important to understand how their behavior differentiates between the two distinct subject areas. For example, studies argue that distribution of IF values in different subject domains

Factor	Answer option	% of authors	
		Medicine	Social sciences
Author charge (Maximum amount in US Dollar)	50	10.0	14.8
	100	3.3	7.4
	200	10.0	22.2
	300	0.0	3.7
	400	0.0	3.7
	500	3.3	25.9
	600	3.3	3.7
	800	0.0	3.7
	1000	16.7	3.7
	1200	3.3	0.0
	1300	0.0	3.7
	1400	3.3	0.0
	1500	10.0	0.0
	1700	3.3	0.0
	1800	3.3	0.0
	2000	10.0	7.4
	2500	13.3	0.0
	3000	3.3	0.0
	above 4000	3.3	0.0
Abstracting & indexing	Academic search elite	4.5	12.0
	ASSIA	0.0	9.6
	Embase	9.1	0.0
	ERIC	0.0	7.2
	Google scholar	47.0	42.2
	Medline	50.0	0.0
	Pro-quest	4.5	20.5
	Pubmed	81.8	9.6
	Scopus	48.5	65.1
	Ulrich	0.0	6.0
	Web of science	48.5	60.2

Table 4.23: Author percentages for answer options (questions 5, 6)

depends on a number of unique characteristics belonging to each domain, and it leads to show large IF value ranges in some subject domains, while others have relatively small ranges (González-Betancor and Dorta-González, 2017; Waltman, 2016). Tables 4.25 and 4.26 show chi-square goodness-of-fit and Mann-Whitney test results respectively, obtained for 15 journal factors of 530 medicine and 417 social sciences journals. The chi-square goodness-of-fit test works similar to the Mann-Whitney U test described in section 4.1.1. This test is used to reveal how well an expected sample

Factor	Answer option	% of authors	
		Medicine	Social sciences
Age of journal	Very recent	5.6	3.8
	Recent	5.6	26.9
	Middle-aged	88.9	84.6
	Old	44.4	42.3
	Very old	38.9	26.9
Processing time	Very short	24.0	21.6
	Short	54.0	51.4
	Average	70.0	77.0
	Long	6.0	8.1
	Very long	2.0	0
IF	Very low	1.7	1.7
	Low	5.0	8.3
	Average	61.7	70.0
	High	75.0	65.0
	Very high	48.3	50.0
Journal's prestige	Very low	1.5	1.1
	Low	7.5	4.5
	Average	61.2	55.1
	High	67.2	77.5
	Very high	70.1	58.4
Publisher's prestige	Very low	2.5	1.5
	Low	2.5	1.5
	Average	66.7	50.0
	High	51.3	78.8
	Very high	38.5	53.0
Acceptance rate	Very low	3.2	4.0
	Low	12.9	16.0
	Average	64.5	72.0
	High	45.2	48.0
	Very high	25.8	22.0
International authorship	0 - 20	7.7	15.9
	21 - 40	35.9	25.4
	41 - 60	51.3	55.6
	61 - 80	30.8	49.2
	81 - 100	17.9	27.0
Number of articles per issue	1 - 10	0	64.3
	11 - 30	50.0	42.9
	31 - 60	64.3	10.7
	61 - 100	21.4	3.6
	Over 100	21.4	3.6
Number issues per year	1 - 2	4.5	34.4
	3 - 5	31.8	56.3
	6 - 8	50.0	34.4
	9 - 12	45.5	25.0
	Over 12	27.3	15.6

Table 4.24: Author percentages for answer options (questions 7-15)

fit with an observed sample, but the test can consider categorical variables instead of continues variables used by the Mann-Whitney U test. Interpretation of the $p-$value is also similar to the description given in section 4.1.1 and this study considers the $p-$value for interpreting the results, while reporting the chi-square value for further studies.

This analysis compares the two subject domains based on the actual values of the journal factors collected in section 3.5.1. Therefore, in addition to examining the differences of journal selection factors in the two subject domains from author's point of view, this study allows to compare the differences of two domains based on the actual metadata values belong to the journals in medicine and social sciences. The chi-square goodness-of-fit test was only applied to factors; presence of affiliation, permanent identifier, online and tracking facility, author charges, and abstracting and indexing services since they were recorded as categorical variables. However, peer-reviewed factor was tested using the Fisher's exact test (Fisher, 1934), because all medicine journals were peer-reviewed, while only 5 social sciences journals were not peer-reviewed. This violated one of the main assumptions of chi-square test since the expected frequency of a category became very small, particularly less than five (Kim, 2017). The Mann-Whitney U test was applied to other factors as they were considered as continuous variables. The analysis attempted to reveal significant differences of these journal factors based their actual values in the two subject domains. This was expected further to confirm whether the journal corpora used by the current study was consistent with or deviated from generally accepted distinctions of journal factors between different subject domains.

In contrast to findings of tables 4.25 and 4.26, the factor values which the authors stated as what they wanted for 15 journal factors can also be differentiated from one subject domain to another. Therefore, the study explored statistically significant differences of the factor values which the authors stated in the second author survey with respect to the two subject domains (see tables 4.27 and 4.28). However, since medicine used different numerical categories for factors; journal's prestige, publisher's

Factor	Chi-square	$p-$value
Presence of an affiliation	32.524	0.000
Permanent article identifier	128.162	0.000
Online submission with tracking	253.687	0.000
No author charges	224.928	0.000
Abstracting & Indexing	246.752	0.000
Peer-reviewed	−*	0.016

* Fisher's test does not provide values for test statistic.

Table 4.25: Comparing actual factor values (categorical) of journals in two corpora

Factor	Mann-Whitney U	$p-$value
Journal's prestige	79032.5	0.000
Publisher's prestige	66444.5	0.000
Processing time	108547.0	0.731
Acceptance rate	107224.0	0.509
Age of journal	104789.0	0.212
IF	98198.0	0.000
Number of issues per year	90941.0	0.000
Number of articles per issue	65287.5	0.000
International authorship	91937.5	0.000

Table 4.26: Comparing actual factor values (continuous) of journals in two corpora

prestige, and IF, than in social sciences, the results for three factors do not necessarily imply the existence or non existence of significant differences of original answers provided by the authors in the two domains. Further, peer-reviewed factor was tested using the Fisher's exact test, as there were only 3 authors from medicine domain who did not consider the factor for publishing.

Factor	Chi-square	$p-$value
Presence of an affiliation	3.498	0.061
Permanent article identifier	0.012	0.913
Online submission with tracking	12.082	0.001
No author charges	10.699	0.001
Abstracting & Indexing	12.422	0.002
Peer-reviewed	−	1.000

Table 4.27: Comparing stated factor values (categorical) by authors in two corpora

Factor	Mann-Whitney U	$p-$value
Journal's prestige	7.0	0.000
Publisher's prestige	0.0	0.000
Processing time	1725.0	0.507
Acceptance rate	754.0	0.828
Age of journal	177.0	0.155
IF	0.0	0.000
Number of issues per year	162.5	0.001
Number of articles per issue	48.5	0.000
International authorship	1068.5	0.262

Table 4.28: Comparing stated factor values (continuous) by authors in two corpora

Answers received for the second author survey were further analyzed to find to what extent the authors were able to achieve the journal factor values they stated in the survey. These results and their implications can be considered as side products of the current research since they do not belong to major objectives. However, the importance of obtained results could assist authors to have a realistic idea about the quality of articles they produce in general. Further, identifying journal factors that give least contribution to reach target journals may be the key to publish in the right journal in future attempts.

The study applied chi-square goodness-of-fit test again to detect statistically significant differences between the factor values the authors stated as what they wanted and the values of corresponding factors observed in journals they eventually published articles in. This test was applied to factors treated as categorical variables. Fisher's exact test was used again for peer-reviewed factor. The Mann-Whitney U test was applied to other factors as they were considered as continuous variables after converting to the numerical scale.

Table 4.29 shows the results for categorical variables, while table 4.30 gives results for continuous variables.

This study compared the similarity between the combination of factor values the authors stated as what they wanted and the combination of factor values of the

Factor	Medicine		Social sciences	
	Chi-square	$p-$value	Chi-square	$p-$value
Presence of an affiliation	1.339	0.247	39.198	0.000
Permanent article identifier	29.741	0.000	19.859	0.000
Online submission with tracking	6.818	0.009	0.040	0.842
No author charges	24.500	0.000	1.752	0.186
Abstracting & Indexing	35.286	0.000	24.973	0.000
Peer-reviewed	–	0.245	–	0.029

Table 4.29: Comparing categorical factor values stated and actual factor values of journals they published

Factor	Medicine		Social sciences	
	M-W*	$p-$value	M-W	$p-$value
Journal's prestige	55	0.000	1165	0.000
Publisher's prestige	34	0.000	116	0.000
Processing time	1071	0.214	2663	0.772
Acceptance rate	321	0.024	782	0.001
Age of journal	65	0.002	281	0.294
IF	138	0.000	60	0.000
Number of issues per year	64	0.000	322	0.009
Number of articles per issue	94	0.853	224	0.006
International authorship	566	0.051	1491	0.016

* Mann-Whitney U statistic

Table 4.30: Comparing continuous factor values stated and actual factor values of journals they published

journals they actually published in. This comparison was anticipated to explain how the overall publishing needs of authors were achieved. The study used equation 3.14 to compute the similarity scores. Figure 4.10 shows the obtained similarity scores for all authors participated in the second author survey. The straight lines in red show the average similarity scores, 0.73841 and 0.66962 for medicine and social sciences respectively.

Mann-Whitney U test was used again to detect statistically significant differences of similarity scores between the two subject domains. The results reported Mann-Whitney U statistic of 2666 with p-value 0.000. In addition, this study checked significant differences of similarity scores between the journal that authors actually published their articles in and the journal they expected (see table 4.31). Similarity

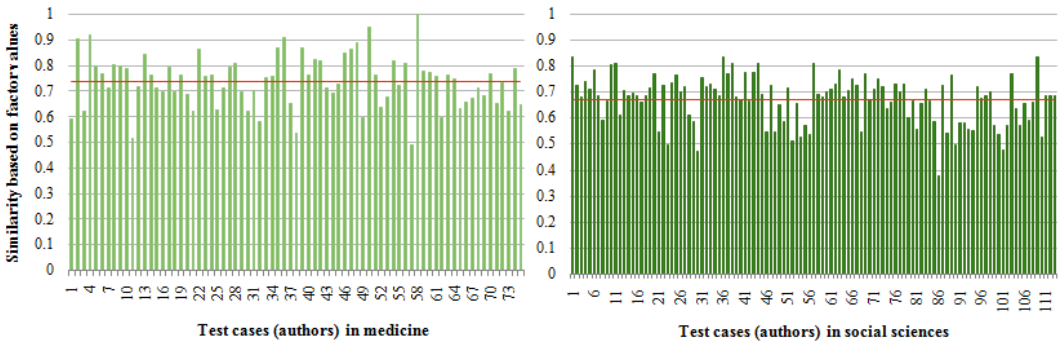


Figure 4.10: Similarity between factor values authors stated and published journals had

score for expected journal was considered as 1 for the Mann-Whitney test, since the study assumed that author's expectations were perfectly fulfilled by the journal, in which the article was expected to be published.

Medicine		Social sciences	
Mann-Whitney U	$p-$value	Mann-Whitney U	$p-$value
37.500	0.000	0	0.000

Table 4.31: $p-$values for similarity between factor values stated and published journals had

4.4 Third author survey: Evaluating the recommender system

Responses received from authors, who replied for the second author survey was used as the sample of the third author survey. Table 4.32 gives brief statistics of third author survey participants.

Third author survey requested authors to indicate the rank according to appropriateness (1-most appropriate to 10-least appropriate, N/A-Not appropriate at all) of suggested 10 journals to publish their given article. The answers received were analyzed along a number of directions to determine the effectiveness of C&K recom-

Sample	Medicine	Social sciences
Total sample	75	113
Emails not delivered	0	0
Effective sample	75	113
Completed survey	61	90
Response rate (approximately)	81.3%	79.6%

Table 4.32: Response rate of third author survey

mender system.

4.4.1 Performance of Content-based and Knowledge-based recommender system

66.2% of journals suggested by the C&K recommender system were indicated as appropriate by authors in medicine, while 58.8% of journals suggested by the C&K recommender system were selected as appropriate by the authors in social sciences domain. This analysis considered each journal, which was assigned a rank as appropriate, independently from the rank.

We used DCG measure to determine how well the C&K recommender system performs comparative to another two gold standards. NDCG was not utilized for this as it needs the number of ordered relevant journals for each input abstract in the corpus. These numbers are required for calculating the IDCG component of the NDCG. On the one hand, these figures are not available for calculations since all corpus journals cannot be ranked practically based on each author's opinion regarding how far the journals are relevant to their article abstract. On the other hand, normalization of DCG is not necessary for comparing performance with a gold standard as the study considered a fixed number of retrieved results (i.e. 10 results).

Equation 3.12 calculated the DCG to determine performance of the C&K recommender system. Since the DCG measure is based on graded relevance of retrieved results similar to the NDCG measure, this problem defined relevance as the reverse

order of the rank assigned by the authors of the third author survey. For instance, a retrieved document with the least rank (i.e. rank 10) attains the least graded relevance (i.e. relevance 1), while a document with the highest rank (i.e. rank 1) receives the highest graded relevance (i.e. relevance 10). Similarly, all subsequent results are mapped their graded relevance based on the reverse order of rank they are assigned. In case of irrelevant results with rank indicated as 'N/A', zero relevance was applied for calculations.

Table 4.33 illustrates an example for assigning graded relevance based on corresponding ranks the results received for all 10 topmost suggestions. Calculation of DCG for the example is also given for further understanding.

Rank order of retrieved results by C&K recommender system: 3, 4, 2, 1, 5, 8, 6, 9, 7, 10. Using equation 3.12 for $p = 10$,

Position of the retrieved result (i)	Rank assigned	Relevance (rel_i)
1	3	8
2	4	7
3	2	9
4	1	10 (highest relevance)
5	5	6
6	8	3
7	6	5
8	9	2
9	7	4
10	10	1 (least relevance)

Table 4.33: Example - defining graded relevance

$$DCG_{10} = rel_1 + \sum_{i=2}^{10} \frac{rel_i}{\log_2 (i+1)}$$

$$DCG_{10} = 8 + \frac{7}{\log_2 3} + \frac{9}{\log_2 4} + \frac{10}{\log_2 5} + \frac{6}{\log_2 6} + \frac{3}{\log_2 7} + \frac{5}{\log_2 8} + \frac{2}{\log_2 9} + \frac{4}{\log_2 10} + \frac{1}{\log_2 11}$$

$DCG_{10} = 28.4$

Finding an appropriate formal gold standard or standard baseline measure for comparing the performance of C&K recommender system is difficult, because there is

no similar system existing at present. Therefore, this study defined two basic gold standards to compare the performance of the current system.

Gold standard 1:

Performance by means of DCG, when the actually retrieved results of C&K recommender system are arranged according to sequential order of most appropriate result at the top of the list to least appropriate result at the bottom of the list. This gold standard is useful to understand the performance of the system as corpus may not include at least 10 appropriate journals for some of the input documents.

Example i: only 3 appropriate results are included in the list retrieved by C&K recommender system.

Rank order of retrieved results: 0, 0, 1, 0, 3, 0, 0, 2, 0, 0

Rank order considered by gold standard 1: 1, 2, 3, 0, 0, 0, 0, 0, 0, 0

Example ii: only 8 appropriate results are included in the list retrieved by C&K recommender system.

Rank order of retrieved results: 4, 2, 1, 6, 3, 5, 0, 8, 7, 0

Rank order considered by gold standard 1: 1, 2, 3, 4, 5, 6, 7, 8, 0, 0

Gold standard 2:

Performance (by means of DCG) of an IR system, which is capable of retrieving appropriate suggestions for all the top 10 results, while arranging them according to the sequential order of most appropriate result at the top of the list to least appropriate result at the bottom of the list. Gold standard 2 deviates from gold standard 1 as gold standard 2 assumes that it retrieves 10 appropriate results sequentially from the most appropriate to the least for every query. Also, the performance of gold standard 1 depends on the performance of C&K recommender system, while gold standard 2 is independent of C&K recommender system. Therefore, gold standard 2 includes two important characteristics. First, it consistently demonstrates a constant performance. Second, it could be considered as the performance of a system, which works at its maximum strength – because all results are relevant/appropriate, while arranging them according to the correct order of appropriateness. Therefore, this

gold standard provides a fairly good maximum performance margin to compare the performance of C&K recommender system.

Example:

Rank order considered by gold standard 2: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Figures 4.11 and 4.12 show the variation of DCG values obtained using actual performance of C&K recommender system, average of C&K recommender system[1], gold standard 1, and gold standard 2 for medicine and social sciences subject domains respectively. Moreover, the approximate DCG values for performance of C&K system and gold standard 1 are tabulated in table 4.34 based on their averages, while the DCG for gold standard 2 is also included for comparisons .

Recommender system	Medicine	Social sciences
Average of C&K recommender system	23.2	21.4
Average of gold standard 1	27.3	26.1
Gold standard 2	29.9	29.9

Table 4.34: DCG for C&K recommender system and gold standards

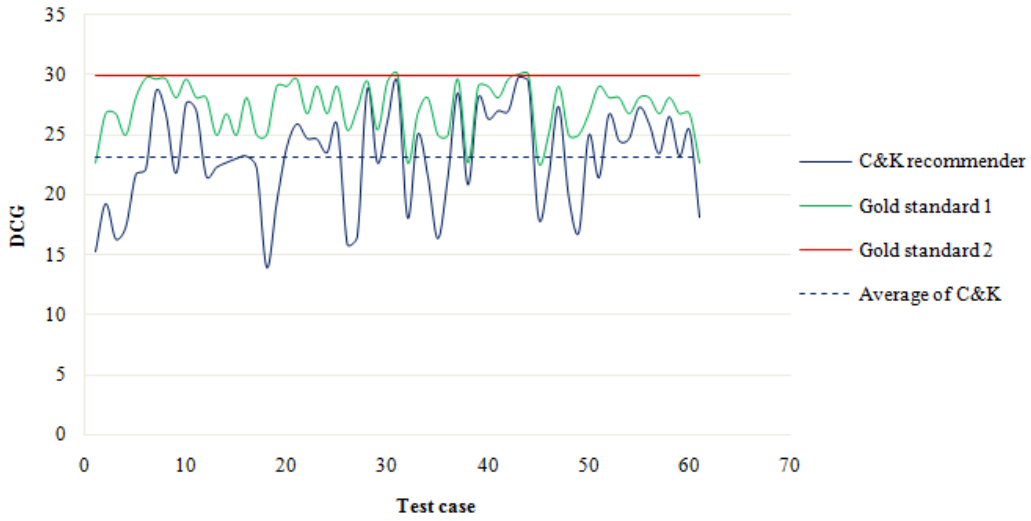


Figure 4.11: Comparing DCG in medicine

This study also determined the number of occurrences C&K recommender system

[1]The average value of the actual performance of C&K recommender system. This average performance is a constant value and easy to use for calculating the performance differences with gold standards.

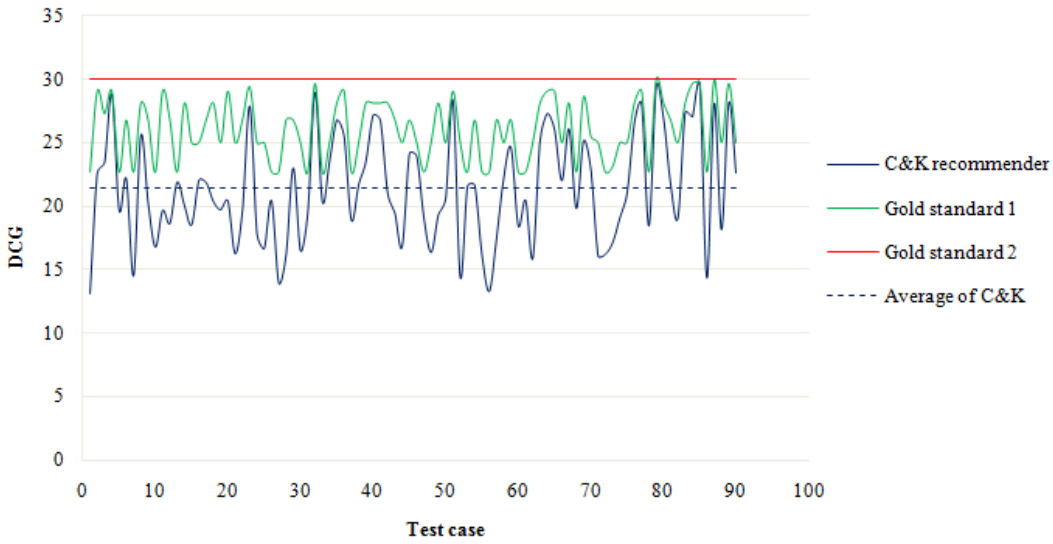


Figure 4.12: Comparing DCG in social sciences

ranked the journal in which the considered article was published within the top 10 suggestions. Table 4.35 shows the number of occurrences of journals, in which the considered articles were published within the top 10 results based on ranks assigned by their authors' opinion.

Rank (author's opinion)	Medicine test cases (total cases = 61)	Social sciences test cases (total cases = 90)
1	29	28
2	5	9
3	4	5
4	5	3
5	-	1
6	1	1
7	1	-

Table 4.35: Number of cases the article published in top 10 results and their ranks assigned by authors

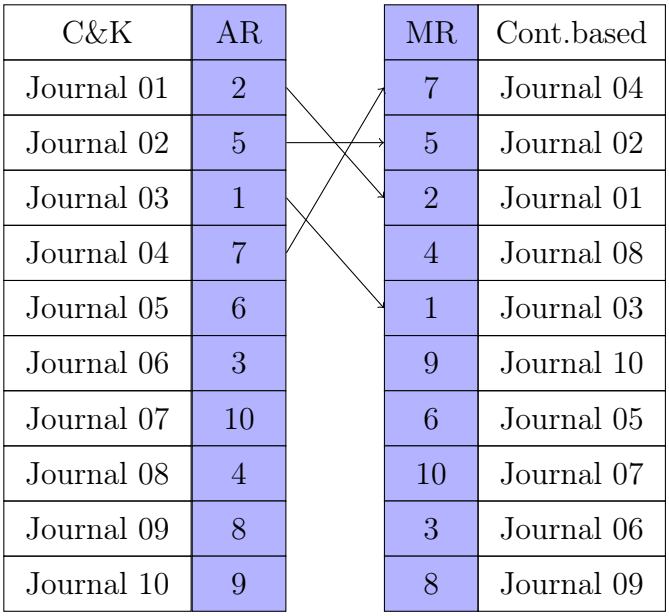
According to the results, 45 retrievals out of 61 in medicine and 47 retrievals out of 90 in social sciences included the journals in which the considered articles were published. The approximate percentages for these figures were 73.8% and 52.2% out of the total retrievals for medicine and social sciences respectively. These figures exceeded the 33% of retrievals reported by the eTBLAST journal finder for the

journal, in which the article was actually published within the top 10 results (Wren et al., 2007).

The number of cases where an author ranked at least one other journal above the journal that he or she published in when the recommender system also ranked the other journal(s) above the actually published journal was 35.5% in medicine and 40.4% in the social sciences. For example, a case − an author assigned rank 1 for *Journal A* and rank 4 for *Journal B* when his or her article was actually published in *Journal B* and the recommender system assigned ranks 2 and 5 for *Journal A* and *Journal B* respectively, was included in this percentage. However, this percentage was applied only for the cases, in which the actually published journal was listed within the first 10 results. These percentages give a notion of how far the C&K recommender system could assist authors to select more appropriate journals if the system existed when they were submitting their articles. To clarify further, 35.5% of authors in medicine whose results were included the actually published journal, had the chance to see more appropriate journal(s) than the journal they actually published in, if the C&K recommender system was available at the time they were deciding the publication venue for the article. Similarly, 40.4% of social sciences authors whose results were included the actually published journal, had the chance to see more appropriate journal(s), if the C&K recommender system was available for them.

4.4.2 C&K recommender system vs. content-based component

Comparing performance between the C&K recommender system and the content-based recommender component is important to find how far the results given by the content-based recommender component was improved by the knowledge-based recommender component. For this comparison, we did not want to conduct a separate survey with the suggestions given by the content-based recommender component



C&K	AR	MR	Cont.based
Journal 01	2	7	Journal 04
Journal 02	5	5	Journal 02
Journal 03	1	2	Journal 01
Journal 04	7	4	Journal 08
Journal 05	6	1	Journal 03
Journal 06	3	9	Journal 10
Journal 07	10	6	Journal 05
Journal 08	4	10	Journal 07
Journal 09	8	3	Journal 06
Journal 10	9	8	Journal 09

Figure 4.13: Corresponding ranks of C&K recommender system and content-based recommenders

alone. Instead, the study used a separate list of journals generated by the content-based component, before sending the list to the knowledge-based component for rearranging their ranks based on authors' criteria of journal selection. Then, the corresponding ranks assigned by the authors for C&K recommender system (i.e. for the third author survey) were mapped to the journals in the list generated by the content-based recommender component.

Example:

Figure 4.13 shows ranks assigned by authors for ordered list generated by C&K recommender system and mapped ranks for ordered list generated by the content-based component. Here, AR and MR denote the rank assigned by authors and mapped rank respectively.

Performance of the two systems were compared based on DCG for the top 10 suggestions in medicine and social sciences (see figures 4.14 and 4.15). Table 4.36 allows to compare the average DCG of results separately for the two systems. Results obtained for Mann-Whitney U test did not report a statistically significant difference of per-

formance between the two systems ($p-$value= 0.882 in medicine and $p-$value= 0.716 in social sciences).

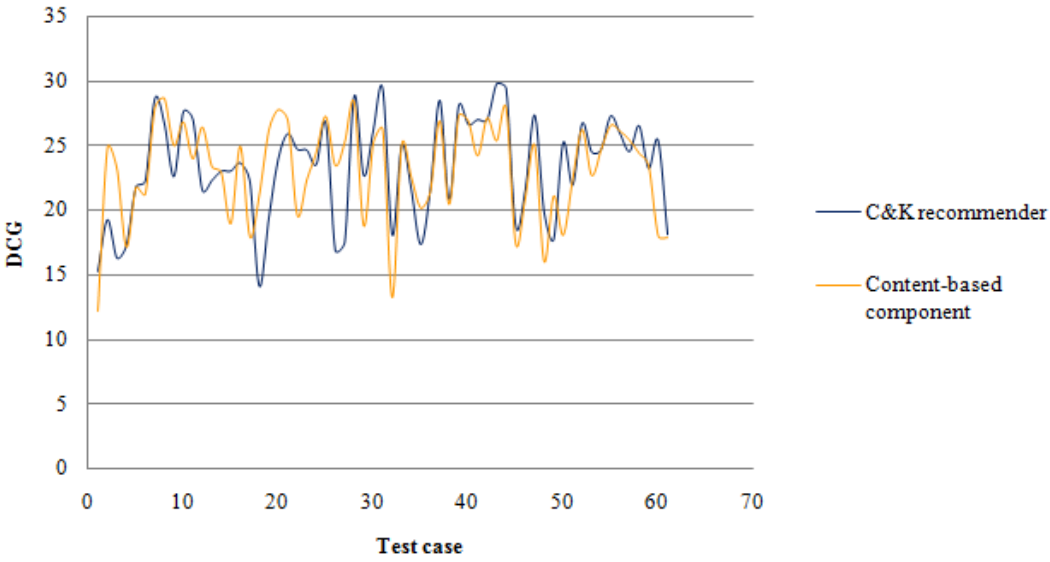


Figure 4.14: Comparing two systems in medicine

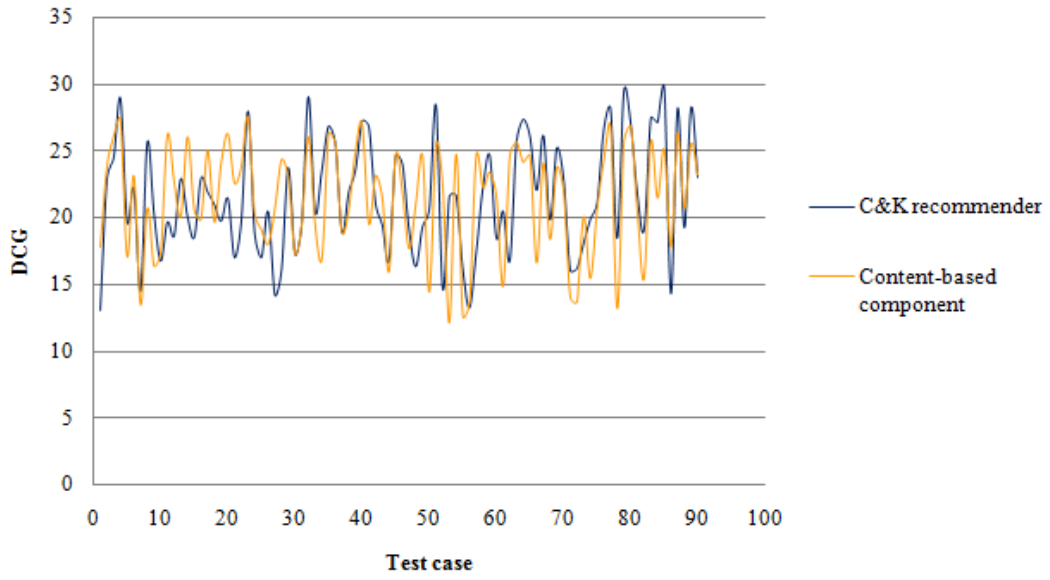


Figure 4.15: Comparing two systems in social sciences

Recommender	Average DCG in medicine	Average DCG in social sciences
C&K	23.2	21.4
Content-based	22.1	20.4

Table 4.36: Average DCG for two systems

This study focused on finding potential reasons for demonstrating statistically insignificant improvement of performance by the C&K recommender system compared to the content-based component alone. These investigations were done along three major arguments.

1. Influence of listing more inappropriate journals close to the bottom of the content-based system, but top and middle places of the C&K recommender system.
2. Influence of inappropriate results generated by the content-based component and advancing them by the knowledge-based component.
3. Influence of the number of factors considered by authors for the recommender systems.

Higher possibility of presenting inappropriate or less appropriate journals close to the tail of the retrieved list possibly leads to a reduction in the performance of C&K recommender system compared to content-based component alone, when considering a relatively long list of suggestions. Content-based recommender is likely to place content wise inappropriate results at the bottom of the list since the suggestions of the system is completely based on the content matching. However, since the C&K recommnder system is based on other factors in addition to the contents, it could bring some bottom ranked journals by the content-based system to the top or middle of the list. Therefore, to check this, the ratio between DCG values for C&K recommender system and content-based recommender alone were plotted as shown in figure 4.17. This ratio was calculated for the top 5 results, in addition to the ratio based on top 10 results. We did not use absolute DCG values or average DCG for comparison as the study targets to compare the performance for top 5 and top 10 results. The absolute DCG values are not appropriate to compare performance for different lengths of retrieved lists.

The study used the same responses given by authors for the third author survey to determine DCG ratio for top 5 suggestions. However, the new lists were limited

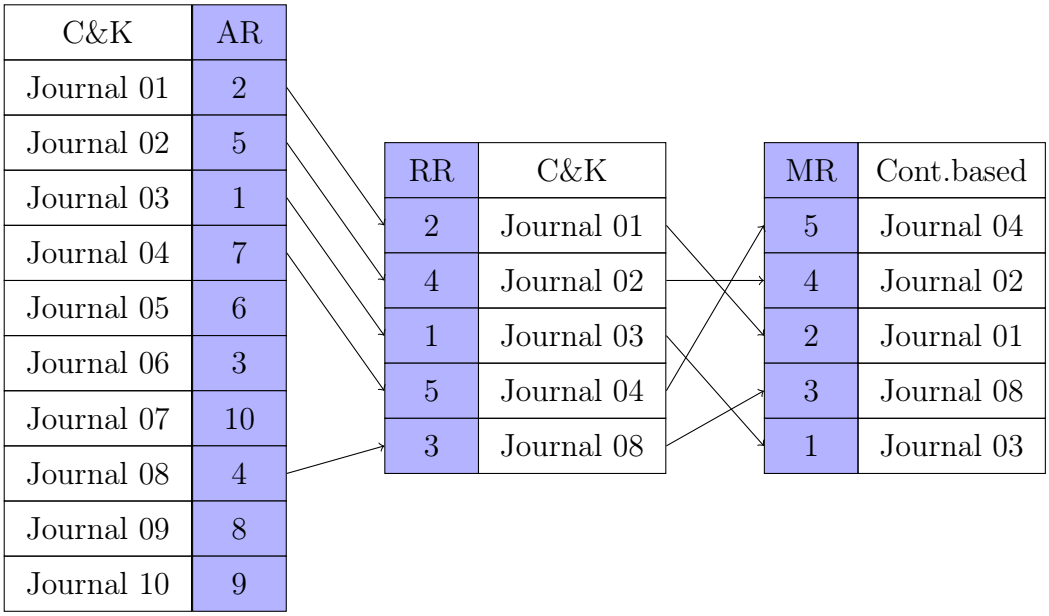


Figure 4.16: Corresponding ranks of C&K recommender system for top 10, C&K recommender system for top 5 and content-based component for top 5

to the 5 topmost results suggested by the content-based recommender component alone. These 5 journals were also included among the 10 suggestions sent to authors for ranking in the third author survey. Therefore, it was possible to re-rank them from 1 to 5 corresponding to the ranks assigned in the list of 10 suggestions.

Example:

Figure 4.16 shows the ranks assigned by authors for ordered list (top 10) generated by C&K recommender system, the re-ranked list after limiting the ordered list generated by the C&K recommender system for only top 5 suggestions and mapped ranks for ordered list generated by the content-based component for only top 5 suggestions respectively. Here, AR, RR and MR denote the rank assigned by authors, re-rank for top 5 and mapped rank respectively.

It is important to note that the list for only 5 topmost suggestions generated by the C&K recommender system includes the same 5 topmost journals suggested by the content-based recommender component since C&K recommender system is able to reorder only the journals suggested by the content-based component in its list

of 5 journals. Moreover, *Journal 01* and *Journal 03*, which are included in the list with 5 suggestions generated by the C&K recommender system receive the same ranks corresponding to the ranks in the list of 10 suggestions. However, *Journal 02*, *Journal 04*, and *Journal 08* in the list generated by C&K recommender system receive different ranks corresponding to the list of 10 suggestions, because all journals in the list of 10 suggestions do not appear in the list of 5 suggestions and it leads to changes in the rank of corresponding journals without disturbing their sequential order.

Figure 4.17 shows ratio between the DCG values obtained by the C&K recommender system and the content-based component alone for all test cases in two subject domains. The blue markers of the figure denote these ratios for the 10 topmost results, while the yellow markers denote the ratios for the case of 5 topmost results. The red line that intersects the $y-$axis at 1 emphasizes the boundary, in which the C&K recommender system and content-based component perform alike. Therefore, the markers above this line point-out the test cases that C&K recommender system outperforms the content-based component. The markers below the red line depict the test cases, in which the content-based component outperforms the C&K recommender system. There are 33 and 50 blue markers above the red line in medicine and the social sciences subject domains respectively. Hence, the C&K recommender system outperforms the content-based component 33 times (54.1%) out of all test cases in medicine domain, while 50 times (55.5%) out of all test cases in the social sciences subject domain when the evaluation considers the 10 topmost results. In contrast, the C&K recommender system outperforms the content-based component 39 times (63.9%) out of all test cases in medicine domain, while 61 times (67.7%) out of all test cases in the social sciences subject domain when the list is limited to top 5 results.

Table 4.37 shows a comparison of performance between C&K recommender system and content-based recommender component alone when considering top 10 suggestions and top 5 suggestions.

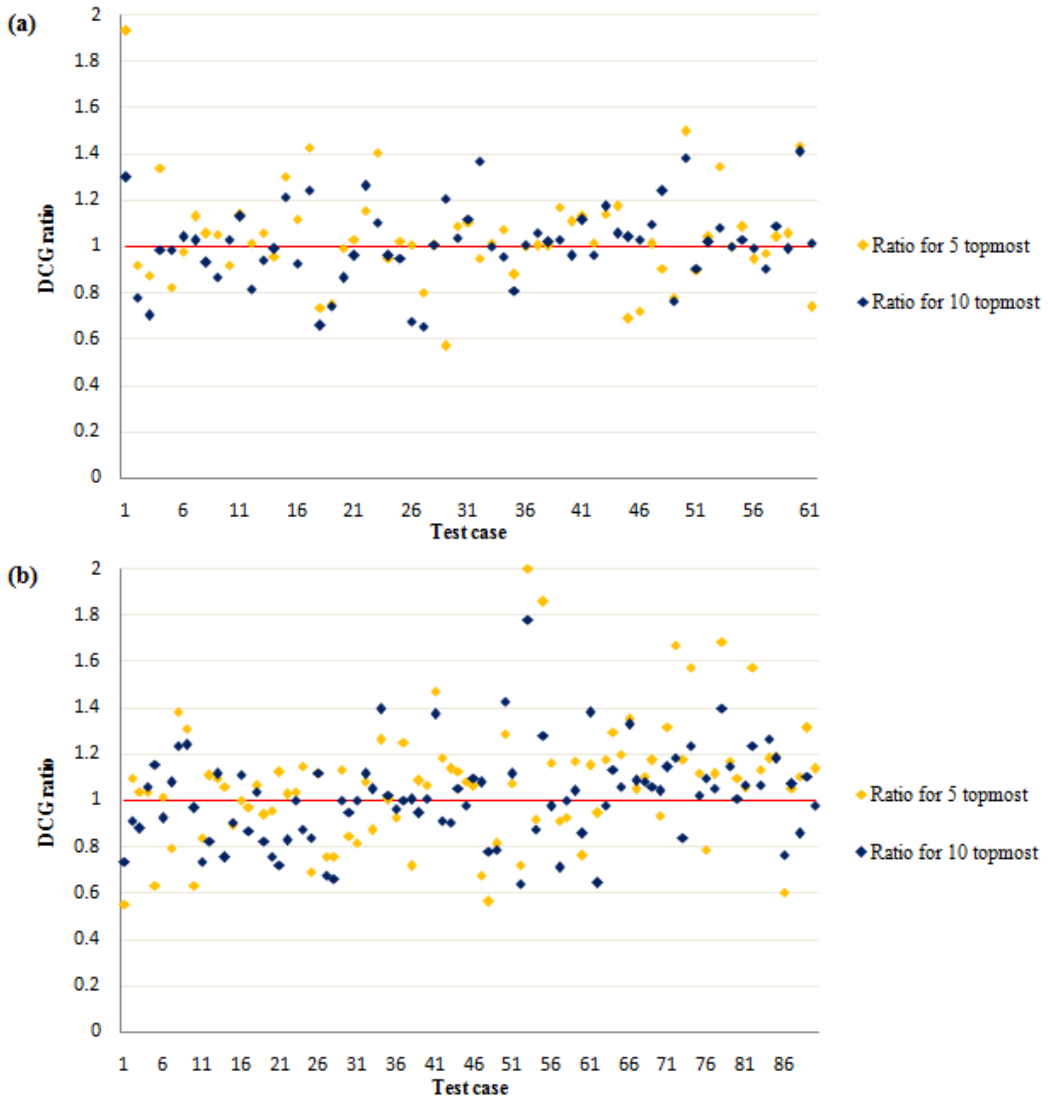


Figure 4.17: DCG ratios for 10 and 5 topmost results of (a) medicine (b) social sciences. Markers above the red line indicate the cases that C&K recommender system outperforms content-based component, while the opposite is indicated by the markers below the red line.

Performance measure	Medicine		Social sciences	
	Top 10	Top 5	Top 10	Top 5
Average of DCG ratio	1.01	1.04	1.02	1.07
% of cases for DCG ratio > 1	54.1	63.9	55.5	67.8

Table 4.37: Performance for top 10 and top 5

The Spearman's correlation analysis was conducted to examine relationship between the ratio of DCG and number of inappropriate suggestions included in the list gen-

erated by the content-based recommender component alone. For this, we considered inappropriate suggestions, which appeared after the fifth topmost result. This approach was expected to check the possibility of improving C&K recommender system over content-based component while limiting the suggestions to 5 topmost results of the content-based recommender component. The obtained results are given in table 4.38. Table 4.37 shows a slight improvement of the C&K recommender system than the content-based component when the retrieved list is limited to 5 journals. However, the results in table 4.38 do not show any strong correlation between the performance ratio and number of inappropriate suggestions after the fifth suggestion of the retrieved list.

	Medicine	Social sciences
Correlation	−0.185	−0.369*

* Correlation is significant at the 0.01 level (2-tailed).

Table 4.38: DCG ratio and inappropriate suggestions

Following evidence shows that increasing performance of the content-based component is a promising approach to improve the C&K recommender system over the content-based system.

1. Figure 4.14 and figure 4.15 imply a potential correlation between the performance of C&K recommender system and the content-based component alone. The Spearman correlation test was conducted and the results revealed a statistically significant and fairly strong positive correlation (see table 4.39) between the performances of the two systems. Therefore, increasing performance of the content-based recommender system is highly likely to increase performance of the C&K recommender system.

2. Considerable number of test cases (46.4% for medicine and 50% for social sciences), in which the C&K recommender system underperformed than the content-based system had the following common characteristic:

Retrieving an inappropriate suggestion for top of the C&K recommender's list. Therefore, Spearman's correlation test was applied to find relationship between the DCG performance ratio of C&K recommender system and the content-based component for the top 10 results and occurrence of an inappropriate suggestion at the top of the list generated by C&K recommender system. Results in table 4.39 show that there exists a statistically significant, fairly strong negative correlation between these two factors in both subject domains. On the one hand, reducing inappropriate suggestions given by the content-based component could also lead to a reduction in delivering inappropriate suggestions by the C&K recommender system to the top of the list. This process is likely to improve the performance of both systems, but could improve the performance of the C&K recommender system by a higher factor than it improves the content-based component. On the other hand, this can be considered as an issue of the knowledge-based recommender component since it cannot avoid delivering inappropriate suggestions to top of the list.

Correlation	Medicine	Social sciences
DCG at 10 of C&K and content-based component	0.663	0.603
DCG ratio at 10 between two systems and inappropriate suggestions at top of the list	−0.549	−0.539

Table 4.39: Correlation of performance for top 10 results and inappropriate suggestions

The number of journal selection factors considered by authors could also impact the performance of C&K recommender system as well as the content-based recommender component. For example, it is reasonable to anticipate higher performance from C&K recommender system than the content-based component, if an author considers a relatively large number of factors. Expecting more contribution from the knowledge-based component while increasing the number of factors is not unusual since this component prioritizes results based on selection factors. Thus, the study applied Spearman correlation test to reveal possible correlation between the number of factors considered by each test case and the DCG performance ratio between C&K

recommender system and the content-based component. The obtained results are given in table 4.40.

Correlation	Medicine	Social sciences
Number of factors and DCG performance ratio at 10 between two systems	−0.103	0.128

Table 4.40: Correlation of DCG performance and number of factors

Although a correlation between the tested variables was anticipated, results reveal that there exists no correlation between the number of factors considered and the performance ratio between the two systems. Normalization property of the algorithm used for knowledge-based recommender component could be the primary reason for this phenomenon. To illustrate, although an author considers large or small number of selection factors, normalizing property of used algorithm may lead to neutralize the effect of the number considered.

Chapter 5

Conclusion

"It is not unscientific to make a guess, although many people who are not in science think it is."

– Richard Feynman: *The Character of Physical Law* (1965), p.165

Selecting a less appropriate journal for publication task leads to disappointment for an author in numerous ways. Chapter 1 of this dissertation specifically focused on these issues and described the importance of finding potential answers to address the problems. The current research primarily aimed to develop a content-based and knowledge-based recommender system to assist authors to select the most appropriate journal outlet to submit their novel articles. Furthermore, the distinct nature of journal selection conversant in different subject domains led to take these differences into account, while developing the recommender system. Therefore, the study selected authors from two different subject domains, medicine and social sciences as the target user groups of the new outcome.

Developing the new journal recommender system is associated with a number of stages. A breakdown of four major stages of the complete process is given below:

1. Identifying and prioritizing author's journal selection criteria in general.

2. Developing a content-based recommender component with an appropriate algorithm.

3. Collecting journal metadata to develop a knowledge-based recommender component.

4. Configuring and evaluating the performance of the hybrid journal recommender system.

Hence, this study made conclusions corresponding to results obtained for each of these four stages.

5.1 Author's criteria of journal selection

The first literature survey conducted to discover the most influential factors of journal selection revealed that an author may consider 16 prominent factors, in addition to another 30 factors only occasionally cited in literature. The survey attempted to discover the weights given by authors in medicine and social sciences for journal selection factors, exploring interesting background information.

European authors reported the highest response rate for the first author survey in both subject domains, while it was minimal from Oceania region. However, it is not possible to decide whether this was caused by the imbalanced geographical proportions of the sample or other factors as we do not have enough location details for the sample.

The survey sample is divided in half between authors having published between 1 to 10 years and authors having published for more than 10 years across both subject domains. Therefore, it is reasonable to consider that the study is not biased towards less or more experienced authors. However, we found evidence to prove that authors in medicine publish more frequently. The percentage of medicine authors, who work as an editor or editorial board member, is also greater than the percentage of social

sciences authors, indicating slightly more expertise and involvement in publishing in the medicine authors' sample (Wijewickrema and Petras, 2017).

The peer-review status of the journal is considered as the most important aspect, which influence the author's journal selection criteria in both subject domains. This factor reported the highest mean among all factors considered. This result is compatible with the findings of a previous research conducted by Dalton (2013). This finding reflects the authors' positive attitude of publishing high quality articles while improving the manuscript before publication. The second most important factor differs across the two subject domains. Medicine authors consider abstracting and indexing databases as the second most important factor of selecting an appropriate publication outlet, but the social sciences authors assign this place to prestige of the considered journal. However, the mean value for importance for a journal's prestige is higher in medicine (4.07) than in social sciences (3.94). In general, medicine authors seem to assign more importance to a majority of factors compared to social sciences. The number of annual subscriptions is the least interesting factor for authors across both subject domains. Consequently, we may conclude that the authors in both domains pay relatively less attention to circulation ability or demand of a journal (Wijewickrema and Petras, 2017).

Table 4.6 shows that four factors – IF, publisher's prestige, A&I services, and online submission with tracking facility, out of all the 16 factors considered are treated in significantly different ways by the authors across the two domains. The IF of considered journal, prestige of publisher, abstracting and indexing databases in which the journal is included, and online submission with tracking facility are significantly more important to medicine authors than to authors in the social sciences. It is obvious that influence of IF can differ widely across distinct subject domains (Dorta-González and Dorta-González, 2013; Scully and Lodge, 2005; Seglen, 1997). Further, Amin and Mabe (2004) have identified major reasons for higher IF values in medicine than in the social sciences. Incidence of relatively higher IFs of medicine journals may lead medicine authors to pay more consideration on the IF than the social sci-

ences authors. The results of the first author survey conclude that the inclusion of a journal in A&I databases is considered as more important by medicine authors than the social sciences authors (Wijewickrema and Petras, 2017). One possible reason behind this could be the visibility and acceptance of prominent A&I databases available for medicine domain. For example, Fangerau (2004) explains the leading role playing by MEDLINE and EMBASE databases as resource discovery tools in medicine literature. In contrast, A&I databases are far less established or standardized in the social sciences. According to Klein and Chiang (2004), prominent services like Social Science Citation Index (SSCI) uses articles to index from multiple subject domains and may not necessarily recognized as dedicated and well established, indexing services in the social science subject domain due to a number of limitations. The first author survey also examined authors' interest towards technical features provided by journals. Availability of an online submission system with tracking facility of a journal is considered more important by authors in medicine than in the social sciences group. Usually, researchers in medicine have to be in contact more with the technology than the researchers in social sciences subject domain. This may reflect the general attitudes of the disciplines to integrate information technology in their research. However, further research on the reasons behind such prioritizing is important to shed more light upon the exact reasons for indicating significantly differing publishing choices as revealed by the survey.

According to section 4.1.1, strong correlations between three pairs of factors were discovered by the survey. Among them, two pairs are valid for both subject domains. The third pair is applicable only for the social sciences domain. Finding correlations between factors is important to examine the nature of influence of one factor on another. Authors can use these findings to construct their journal selection criteria more effectively by including or excluding the correlated factors. However, out of these three, only one correlation can be explained with general understanding of publication procedure. Importance of the number of issues a journal publishes and the number of articles it publishes per year are correlated, because the number of articles highly depend on the number of issues for most cases (Wijewickrema and

Petras, 2017). Further research is needed to explain the discovered relationships between author contributions from different countries and the existence of a persistent identifier or the number of articles published per year and the online submission system with tracking facility.

The questions probed authors' awareness of journal recommender systems, which could be used as supportive tools to make the journal selection process smoother – which led to the discovery that only less than half (35%) of the authors in both subject domains knew of the existence of such assisting tools. Majority of authors, who are aware of the existence of journal recommender systems use them either often or infrequently to select an appropriate publication outlet. Therefore, academic institutions, publishers and other research organizations should take the opportunity to consider these indications seriously, to implement necessary programmes to make researchers aware of the resources they can utilize in publishing. Moreover, despite the neutral answers received, a considerable percentage of authors (67% in medicine and 55% in social sciences) in both subject domains, believe that journal recommender systems are helpful to select an appropriate journal to submit an article. On the one hand, this finding elaborates authors' positive attitudes towards their usefulness in publication process, while on the other hand, it implies the importance of the current study as it aims to develop a novel recommender system for journals.

A careful inspection of table 4.12 leads to a meaningful interpretation for the three components which summarize 16 journal factors. Examining shared characteristics of each factor under common components emphasized that component 1 is associated with factors, which refer to a journal's scientific reputation. Factors in component 2 reflect performance or production issues of a journal, while aspects in component 3 relate to the reliability of and the demand for a journal. It is not surprising that the same factor can belong to two different components due to their intrinsic characteristics. For instance, the number of journal issues per year can represent the performance of the journal (with respect to its output capacity), but also its reliability and demand (with respect to its publishing strength and readers' interest). Despite

assigning different importance levels to journal factors, both groups of authors recognize the 16 factors under the same components. Average importance assigned by authors across both subject domains for the three components shows that authors are more concerned about the reputation of a journal than about its performance ability (see figure 5.1). Publishing in a well-known and reputable journal is more important than getting an article published fast. Reliability and demand gains the least attention out of the three components. This compels one to hypothesize about the authors' limited ability to view a publication as a social contribution (Wijewickrema and Petras, 2017). For example, a permanent identifier of an article ensures persistent access to the publication for a longer period. Guaranteeing this reliability is in fact a social contribution as the article reaches a few more generations ahead. Increasing annual subscribers can also be seen as a social contribution as it supports increasing the reach of findings of research to a wider community.

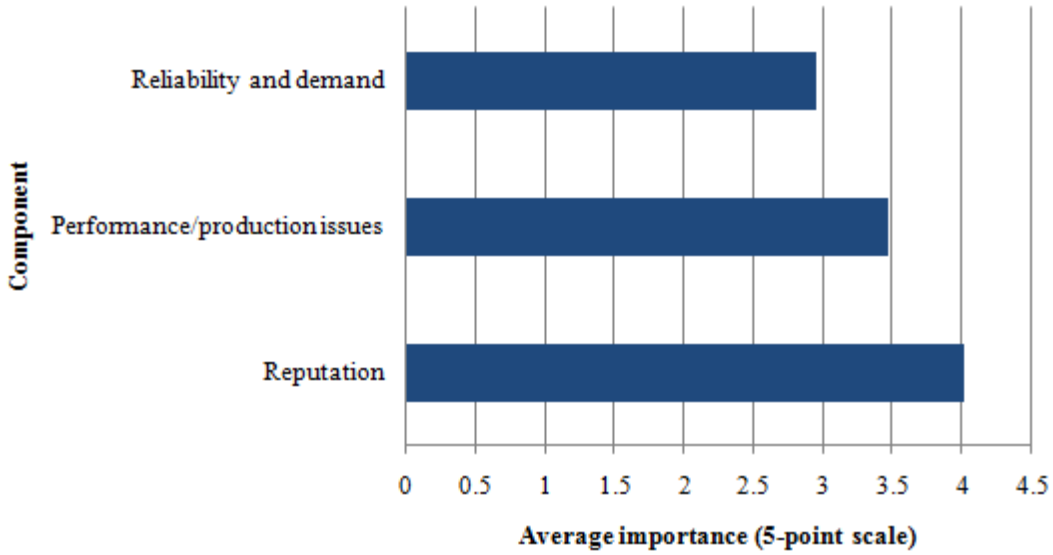


Figure 5.1: Average importance of components

5.2 An algorithm for recommender system

A rapid increase in the number of journals can lead authors to choose a less appropriate journal outlet to submit a manuscript while ignoring many fitting journals. As a solution, automatic recommender systems are proposed to assist authors. However, their performance depends on a number of factors such as the subject of corpus, the algorithm comparing the test document with corpus, and the nature of the corpus documents. Part of the current research addresses this issue by developing a novel content-based journal recommender component for two radically different OA subject corpora while selecting the most appropriate algorithm for comparing each corpus with test documents.

Out of all five algorithms, BM25 can be considered as the most appropriate similarity measure across both subject domains. The unigram language measure shows the lowest performance. Further, tables 4.15 and 4.16 reveal that there exists a significant difference in the average NDCG scores obtained by each pair of the five algorithms in each sub-discipline of the two subject domains, except for unigram language model and SVM in social sciences. Therefore, it is reasonable to select BM25 as the best text similarity algorithm out of these five to implement in a content-based recommender system for the medicine and social sciences OA journals. Moreover, since the medicine and social sciences subjects are radically different domains, one could hypothesizes the validity of the argument for some other subject domains. Figure 5.2 gives the summary of average NDCG values scored by the five algorithms in each subject domain. This clearly indicates that all measures follow similar behavior patterns across both subject domains. However, only BM25 shows a slight improvement of average NDCG for social sciences compared to medicine. Usually, the vocabulary used in medicine is more technical and comprises more specific terms than in other subjects including the social sciences domain. Perhaps, this resulted in the finding of more relevant journals in medicine than in the social sciences subject domain and consequently a better performance in medicine compared to the social sciences subject domain (Wijewickrema et al., 2019). Tables 4.17 and 4.18 show that

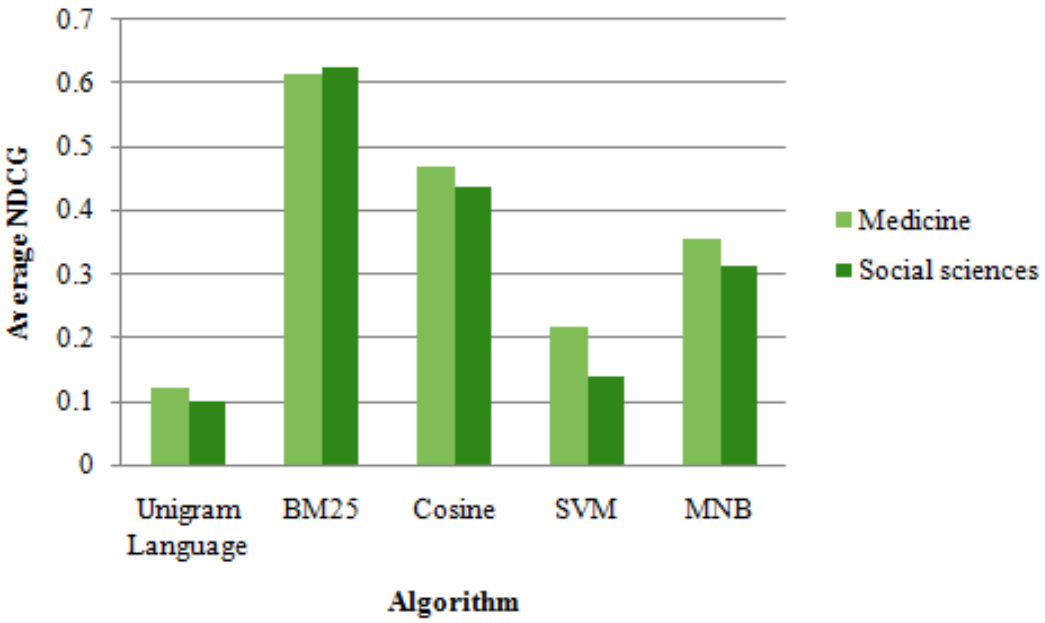


Figure 5.2: Average NDCG of algorithms in the two subject domains

there exists a moderate and positive correlation between each pair of BM25, cosine, and MNB algorithms in sub-disciplines of each subject domain. However, these three algorithms have small or negligible correlations with the unigram language and SVM algorithms. Features such as using term frequency-inverse document frequency based components and fairly similar inverse document frequency expressions may be the reason behind this relatively strong correlations between the BM25 and cosine similarity (Wijewickrema et al., 2019), but these characteristics do not hold by the MNB. The length normalization of corpus documents in BM25 may be the reason for the better performance over the cosine similarity and MNB. On the one hand, equation (3.3) yields that the BM25 improves performance for corpora with large average document lengths. On the other hand, figure 4.7 shows that the social sciences corpus has higher average corpus document lengths than in the medicine subject corpus. This reveals the reason for higher average NDCG scores with BM25 in the social sciences domain than in medicine (see figure 5.2). Overall, the observations imply that the unigram language model and SVM deviate considerably from the BM25, cosine, and MNB algorithms while the latter three follow relatively similar patterns of behavior.

The qualities of the corpus may also impact the results of the recommender system. The outcome of this research produces evidence to declare that cosine similarity advances its performance in sub-disciplines and subject domains with higher technical vocabulary while demonstrating an approximately similar distribution to BM25 algorithm in the medicine subject domain (Wijewickrema et al., 2019):

1. Figure 4.6 illustrates that cosine similarity outperforms all other four algorithms for the 'Technology' sub-discipline in the social sciences subject domain. In general, according to the nature of the 'Technology' sub-discipline, it may comprise comparatively more specific vocabulary than in most of the other sub-disciplines.

2. Figure 5.2 visualizes that cosine similarity attains higher average NDCG score for medicine and performs better in medicine than in the social sciences subject domain. Medicine articles of the training corpus are likely to have a relatively higher amount of technical terms than the articles of social sciences training corpus.

3. The histograms obtained for NDCG of all test documents in both subject domains (see figure 5.3) indicate that BM25 and cosine similarity have approximately similar distributions in medicine while the two distributions in the social sciences subject domain are skewed in opposite directions. This is an evidence for the similar behaviors of cosine similarity and BM25 in a subject domain like medicine, which has rich technical vocabulary.

The current study also examined the performance of five algorithms against the average article lengths of the journals belonging to each sub-discipline in the two domains. Figure 4.7 illustrates that the average article lengths of all sub-disciplines are approximately similar for the medicine subject domain while the average article lengths of the social sciences sub-disciplines vary from one to another. Moreover, the average article lengths in the medicine subject domain are significantly lower than in the social sciences subject domain. One of the important observations is given by the

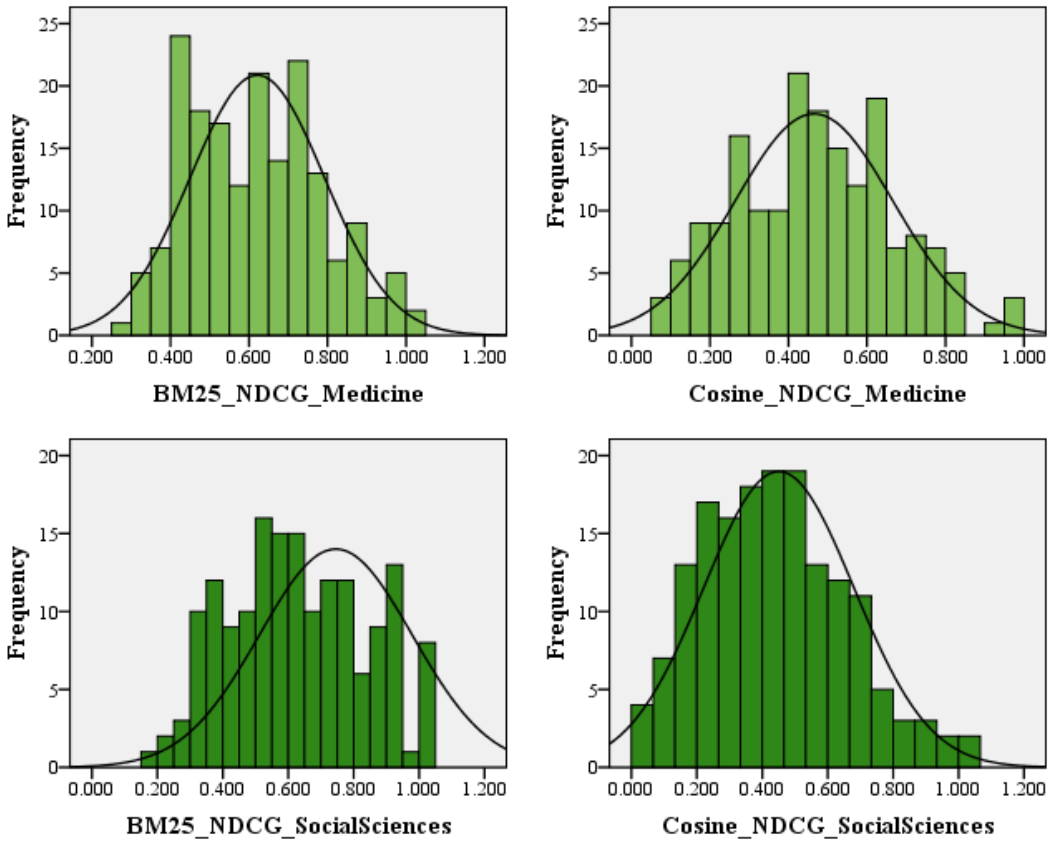


Figure 5.3: Distributions of NDCG of BM25 and cosine similarity. Algorithms have approximately similar distributions in medicine than in the social sciences.

correlation results (see table 4.19) between the five algorithms and the average document lengths of sub-disciplines. They negatively correlate in the medicine subject domain. However, there are only two moderate correlations – between the average article lengths of medicine and cosine similarity, in addition to SVM. As a result, we can conclude that documents with shorter lengths in medicine give better results for the cosine and SVM algorithms. A potential reason for the low performance for lengthy training documents could be the noise caused by relatively large number of less significant terms in them. Gatta et al. (2014) point out the phenomenon of lowering the classification performance due to the presence of less significant terms in large training corpora.

The number of journals that represent the sub-disciplines in the training corpus may also influence the recommender system performance in addition to the length of the

corpus documents. Table 4.20 illustrates the existence of moderate and positive correlation between the performance of the MNB algorithm and the number of journals belonging to the sub-disciplines of the social sciences corpus. Performance of the cosine similarity also shows slightly low correlation in the medicine subject domain, but it is positive and moderate with the number of journals belonging to the sub-disciplines of the social sciences subject domain. These two observations imply the possibility of improving performance of MNB and cosine similarity by increasing the number of training journals in the social sciences corpus. Moreover, the ability of improving the performance parallel to increasing the number of corpus journals could be relatively higher for the MNB algorithm than for the cosine similarity. However, as emphasized by Gatta et al. (2014), continuously extending the size of the training corpus will not significantly affect the performance of the similarity algorithms, because there could be an upper bound of the training corpus size that gives the optimum results. This argument is further supported by an experiment done by Banko and Brill (2001). The upper bound that offers the optimum performance for the social science corpus of the current research can be found empirically by a future research. In contrast, the number of corpus journals gives a low or negligible correlation with the performance of all five algorithms in the medicine subject domain. Accordingly, changing the number of journals in the medicine subject corpus may influence the performance of none of the five algorithms.

5.3 Journal metadata and author's expectations

Basic results of the second author survey indicate that the highest percentage of authors across both subject domains want to select a peer-reviewed journal to submit their articles. This observation aligns with the results of the first author survey, because it reveals that authors in both domains give their highest attention to peer-reviewed factor when selecting an appropriate journal. The factor which receives the attention of the least number of authors in medicine is reported as the average

number of articles per issue, while this place was attained by the factor — journal's age, in the social sciences. The latter result is consistent with the results of the first author survey as the factor – journal's age reports least importance in the social sciences domain. Another, primary, but significant finding is that the majority of medicine authors, those who do not necessarily consider journals which are only free of charges are in a position of bearing higher author charges. In contrast, journals with relatively low author charges are preferred by the majority of authors in the social sciences. This is supported by two previous studies (Solomon and Björk, 2016; Solomon and Björk, 2012) since they found that payments made for APCs within a year for medicine related journals exceeds the cost of APCs for social sciences in the same year. Availability of more financial support for publication costs in medicine than for authors in social sciences and higher publication charges of medicine journals compared to social sciences journals could be two potential reasons to explain this trend, though we need additional information to confirm it. Moreover, the results of the second author survey conclude that target journal's inclusion in Pubmed database is of concern to the highest percentage of authors in medicine, while the highest percentage of authors in social sciences domain consider Scopus database. This finding slightly contradicts with the A&I databases ranking criteria constructed in table 3.4, because WoS is indicated as an important database by the highest percentage of literature concerned. Therefore, a future study can be considered to revise the ranking criteria of A&I services based on author's expectations in the distinct subject domains.

The study used the actual metadata of the journals to examine the differences between the two journal corpora further. The nature of journals in the two subject domains show statistically significant differences for most factors, except processing time, acceptance rate and journal's age. Thus, the latter three factors can be considered as non-volatile factors for distinct subject domains like medicine and social sciences. This supports the argument that the medicine and social sciences journals indexed in DOAJ started their publication in more or less similar years and continue publication of new journals parallel to each other. In addition, similarities

of management of publication process between the two domains can be observed to some extent due to non significant differences of factors like processing time and acceptance rates.

According to table 4.24, the highest percentage of authors in both subject domains select highly reputed, middle-aged journals with average processing time, average acceptance rates, and 41-60% of international authorship per issue. However, the highest percentage of medicine authors prefer high IF journals, while an average IF is sufficient for most social sciences authors. An average publisher's prestige is expected by most medicine authors, but most social sciences authors want to submit their articles to journals with higher prestige publishers. 6-8 issues per year are selected by majority of authors in medicine, but this amount is only 3-5 for most social sciences authors. A substantial difference between values of factors in two subject domains is noticed for the number of articles per issue. This is 31-60 for most medicine authors. However, only 1-10 articles per issue are selected by the majority of authors in social sciences to submit their articles. Therefore, the highest percentage of medicine and social sciences authors who participated in the second author survey expected approximately similar standards for most factors considered. This result aligns with the findings of the first author survey since it revealed that only four journal selection factors are considered significantly differently by the authors in two domains.

The journal factor values which the authors stated as what they expected from journals do not show statistically significant differences between the two subject domains for factors – presence of an affiliation, having a permanent article identifier, peer-review status, processing time, acceptance rate, age, and international authorship. Further, one can observe that the actual journal metadata values of three factors – processing time, acceptance rate, and journal's age do not significantly differ between the two subject domains (see table 4.26). However, the actual journal metadata values of other four factors – presence of an affiliation, having a permanent article identifier, peer-review status, and international authorship differ significantly

between the two subject domains. This leads to conclude that existence of significant differences of the actual factor values of latter four factors between the journals of two subject domains do not influence to differentiate author's expected values of these four factors significantly in medicine and social sciences.

There exist statistically significant differences between author's expectations regarding factor values they stated and the factor values of the journals they actually published in. The factors – permanent article identifier, A&I databases, journal's and publisher's prestige, acceptance rate, IF, and issues per year show significant differences in both subject domains. This means the authors in medicine and social sciences do not succeed in publishing articles in a journal, where the expected factor standards are available for above seven factors. Online submission with tracking facility, no author charges, and age of journal show significant differences between author expectations and published journals only for medicine domain. Author's expectations in social sciences are not fulfilled by the published journals regarding factors – presence of an affiliation, peer-review status, number of articles per issue, and international authorship.

The study also determined a composite score for overall similarity of 15 journal factors between author's expected values and the factor values really existing in published journals. According to findings, the overall standard expected by authors from published journals deviates significantly from the standard they really achieved. Moreover, this failure is common for authors in both subject domains. However, the higher average similarity between author's stated values and journal's actual values in medicine than in the social sciences concludes that medicine authors are more likely to achieve their publication needs from journals than social sciences authors. This argument can be used as an initiation to explore more information on why the medicine authors are slightly ahead of social sciences authors in terms of fulfilling their publication tasks. Availability of more publication opportunities, particularly relatively higher amount of OA journal outlets in medicine could be one close reason for this. For example, the current research found that there were 1154 medicine and

658 social sciences journals in DOAJ during the period of late 2016 to early 2017.

5.4 Evaluating C&K recommender system

The third author survey aimed mainly to evaluate the performance of the journal recommender system developed using a content-based and knowledge-based recommender components. Majority of journals suggested by the C&K recommender system to publish authors' articles were accepted as appropriate by the corresponding authors. C&K recommender system suggests approximately 66.2% of appropriate journals for input articles in medicine domain. However, the performance of the system relatively reduced for input articles in social sciences domain. Only 58.8% of suggestions are selected as appropriate by the authors in social sciences. Therefore, despite the performance of ranking, C&K recommender system tends to give slightly higher number of appropriate suggestions in medicine domain than in the domain of social sciences. Inclusion of relatively less number of appropriate journals in the social sciences corpus corresponding to input articles can be envisaged as a key reason for this underperformance in the domain of social sciences. This assumption is further supported by the size of medicine corpus, because it includes considerably more journals than in the corpus of social sciences. Inclusion of more journals in a corpus is likely to retrieve more number of relevant results for top n suggestions than when it uses a corpus with less number of journals to retrieve the same number of suggestions. Therefore, increasing the size of social sciences corpus could lead to suggest more number of appropriate journals for its authors. In addition, the difficulty of disambiguation of the social sciences articles by the content-based recommender component may lead to demonstrate a lower performance in the social sciences corpus than in the medicine.

This study compared the performance of newly developed C&K recommender system with two other gold standards. The average performance of C&K recommender system makes 22.4% and 28.4% performance difference in medicine and social sciences

domains respectively with a recommender system which gives appropriate results for all top 10 suggestions, while ranking them from most appropriate to least (i.e. gold standard 2). The performance differences between C&K recommender system when it demonstrates its average performance and when it ranks all retrieved results (both appropriate and inappropriate) from most appropriate to least appropriate (i.e. gold standard 1) are 15% for medicine and 18% for social sciences. For both cases, the average performance of C&K recommender system shows performance loss against gold standard 1 and gold standard 2.

Finally, the study evaluated how far the results given by the content-based recommender component is improved by merging with the knowledge-based recommender component. Figure 5.5 and figure 5.6 show the output windows of the two recommender systems given for a same article abstract (see figure 5.4) authored in medicine domain. For this example, C&K recommender system reported 26.9 of DCG, while content-based recommender component reported 24.1 of DCG.

C&K recommender system outperforms the content-based component slightly in terms of both number of test cases and average DCG. However, making the retrieval list half of the original size shows more enhancement of performance of the C&K recommender system than the content-based recommender component alone. Results of the current research are not sufficient to reveal the exact reasons for slight improvement of C&K recommender system with a shorter list of results given in table 4.37, but proves it is not influenced by the presence of inappropriate suggestions close to the bottom of the list suggested by the content-based recommender component (see table 4.38).

Presence of inappropriate suggestions at top of the list generated by the C&K recommender system demonstrates a negative correlation with outperforming of C&K recommender system over the content-based component. This evidence implies two possible approaches for improving C&K recommender system. First, minimizing the number of inappropriate results generated by the content-based component could also minimize the possibility of delivering these inappropriate results to top of the

The effect of aging on pacing strategies of cross-country skiers and the role of performance level

The participation of master cross-country (XC) skiers in training and competition has increased during the last decades; however, little is known yet about whether these athletes differ from their younger counterparts in aspects of performance such as pacing. Therefore, the aim of the present study was to examine the combined effect of age and performance (race time) on pacing in cross-country (XC) skiing. We analyzed all finishers ($n = 79,722$) in 'Vasaloppet' from 2012 to 2017, the largest cross-country skiing race in the world, classified according to their race time into 10 groups: $3-4$ h, $4-5$ h, ..., $12-13$ h. A trivial main effect of sex on total pace range was observed ($p < 0.001, \eta^2 = 0.002$), where women ($44.1 \pm 10.2\%$) had larger total pace range than men ($40.9 \pm 11.8\%$). A large main effect of performance group on total pace range was shown ($p < 0.001, \eta^2 = 0.160$), where the smallest total pace range was $21.8 \pm 1.9\%$ ($3-4$ h group) and the largest $50.1 \pm 9.9\%$ ($10-11$ h group). A trivial sex×performance group interaction on total pace range was found ($p < 0.001, \eta^2 = 0.001$) with the largest sex difference in pacing shown in $9-10$ h group. A trivial and small main effect of age was found in women ($p < 0.001, \eta^2 = 0.005$) and men ($p < 0.001, \eta^2 = 0.011$), respectively, where the masters had smaller total pace range than their younger counterparts. A trivial age group×performance group interaction on total pace range was observed in both women ($p < 0.001, \eta^2 = 0.008$) and men ($p < 0.001, \eta^2 = 0.006$) with smaller differences among age groups in the faster performance groups. In summary, master XC skiers adopted a relatively even pacing independently from their race time and the differences in pacing from the younger XC skiers were more pronounced in the slower masters. These findings suggest that exercise attenuates the decline of performance in master XC skiers as shown by the similar pacing strategies between fast master XC skiers and their younger counterparts.

Keywords: Age; Endurance exercise; Gerontology; Sport performance; Winter sport

Figure 5.4: Input abstract from journal "Eur Rev Aging Phys Act."

list generated by the C&K recommender system. Second, improving the knowledge-based recommender component to avoid delivering inappropriate results to top of the list.

The number of journal selection factors considered by authors does not influence outperforming of C&K recommender system over the content-based component. However, this can be caused by an internal property of the algorithm used to implement the knowledge-based recommender component. Equation 3.14 explains its ability

```
List of Appropriate Journals:

[1] European Review of Aging and Physical Activity
[2] Biomedical Human Kinetics
[3] Sports Medicine, Arthroscopy, Rehabilitation, Therapy and Technology
[4] Health Economics Review
[5] Physical Education of Students
[6] Journal of Cancer Epidemiology
[7] Revista Andaluza de Medicina del Deporte
[8] Fisioterapia em Movimento
[9] Sports Medicine - Open
[10]Journal of Fitness Research
```

Figure 5.5: Output of C&K recommender system

```
List of Appropriate Journals:

[1] Sports Medicine - Open
[2] Biomedical Human Kinetics
[3] Sports Medicine, Arthroscopy, Rehabilitation, Therapy and Technology
[4] Journal of Fitness Research
[5] Fisioterapia em Movimento
[6] European Review of Aging and Physical Activity
[7] Physical Education of Students
[8] Health Economics Review
[9] Revista Andaluza de Medicina del Deporte
[10]Journal of Cancer Epidemiology
```

Figure 5.6: Output of content-based recommender component

of minimizing influence of the number of factors considered for performance changes due to normalizing property. Accordingly, order of the list of results generated by the C&K recommender system may not really be influenced by the number of factors considered. Therefore, authors' willingness to consider the content similarities as more important compared to bibliometric factors of journals could still be a strong reason for not demonstrating a significant advancement of the C&K recommender system than the content-based component.

In addition, recalling equation 3.14, the knowledge-based component uses weights for importance assigned by the authors for each factor when deciding the most appropriate journal. These average weights were collected by the first author survey and utilized in the algorithm for knowledge-based component. However, for some

authors, these average weights would not be appropriate. For example, an author may consider the factor 'impact factor' as far more important than that considered by the average authors. In such cases, a considerably higher weight has to be assigned for the impact factor. This dynamic nature of author's interests could lead to a reduction in the performance of the knowledge-based component. Therefore, this incidence can also be considered as a reason for insignificant advancement of the C&K recommender system. However, allowing individual authors to assign factor weights based on their own opinions can be experimented to confirm its influence. Less precise inputs to the knowledge-based component of C&K recommender system could also lead to demonstrate insignificant advancement. If the second author survey had allowed authors to choose numerical values for answer options instead of nominal categorical options like small, average, high, and so on, perhaps the study would be able to obtain more precise results since there could be differences between how the options are mapped from words to numerals by respondents and the current research. However, this study selected nominal categorical options since specifying the numerical values of factors could be difficult for respondents to answer. Finally, inadequate numbers of appropriate journals in each corpus could also lead to suggestions of inappropriate journals by the content-based system. Presence of less than 10 appropriate journals corresponding to a given article abstract in the corpus reduces the system performance for the 10 topmost results. Therefore, in addition to upgrading the content-based component, populating the corpus with journals of more and diverse range of sub-disciplines may cause to enhance the performance of C&K recommender system over content-based component. However, the ability of improving the corpus is restricted to some extent as the corpus is limited to journals in DOAJ.

5.5 Significance of the study

Findings of the first author survey are more important for less experienced researchers in both medicine and social sciences subject domains to distinguish the most relevant publication factors. Since factors' weights can be considered as a reflection of publishing trends of a domain and publication requirements that authors need to prioritize in their domain, fresh authors will be informed what they have to target from the beginning of their career. This will ultimately help young researchers to achieve their career goals such as promotions, institutional rewards and research grants within a shorter period of time. Publication trends of distinct subject domains are important for journal editors and publishers too. Prioritizing the journal factors and adjusting them according to author's needs will attract more authorship to a journal. For instance, editors as well as publishers across both subject domains can be motivated to upgrade a non-peer-reviewed journal to the status of peer-reviewed, because of the significantly higher importance received by the peer-review factor. Results of the first author survey can be combined with the results of the second author survey to make a journal more author friendly. For example, since the factor – processing time of a journal gains relatively higher weight in social sciences and most authors of this domain prefer to select a journal with an average processing time, editors and publishers can give priority to adjusting journal's usual processing time to neither too short to assure a comprehensive review process, nor too long to avoid an obsolete publication. Side products of this study like identification of significant differences of factors between the two subject domains and studying correlations between journal factors will contribute a lot for the development of bibliometrics research. Further, potential usefulness of journal recommender systems revealed by the authors will encourage information systems development researchers to overcome the negative aspects of existing systems by replacing them with enhanced tools.

The results of the second author survey will assist authors to have a rational idea about the quality of research and articles they produce in general. Further, identifying journal factors that contribute inadequately to publishing in the target journal

and studying potential reasons for these failures are important to getting to publish in the right journal next time. Implications are imperative for academic institutions to design and prioritize their career development programs. Correct identification of journal characteristics that authors failed to achieve can be used to direct employees to appropriate training programs about publishing. Lack of resources apart from training could also cause the distance between authors' target journals and achieved journals. For instance, an author may fail to publish in the target journal, because of financial restrictions. Therefore, recognizing the necessity for resources to improve employees' ability to reach target journals can be considered as a beneficial outcome of the study.

C&K recommender system developed by this research mainly aims to minimize the problems faced by authors when selecting an appropriate journal outlet. Thus, avoiding lengthy time span for finding and short-listing several journal options will support them to save more time for scholarly works. The recommender system will aid editors, publishers, and recommender system developers additionally to authors. Journal editors can utilize the current recommender system to check the suitability of a fresh submission to their journal before reviewing them further. The retrieved list of journals generated by the recommender system will direct editors to compare the scope of the manuscript with their journal. Publishers can apply journal recommender systems to motivate their authorship to choose the most suitable journals from their databases. Whenever the publisher's journal database does not comprise medicine or social sciences subject domains, they still can replace the existing corpus with their own subject domain as the current system is adjustable. Thus, customization is possible according to a publisher's needs of the journal database. Directories like DOAJ can widen author services by implementing a journal recommender system to facilitate authors to find a suitable OA journal via cross searching of several OA databases. Professional article editing services, which suggest appropriate journals based on their expert knowledge can apply the new journal recommender system as a filter to refine a long list of potential journals for a given article. In addition, the recommender system developers can use the outcome of this research to enhance ex-

isting recommender systems or to develop novel recommender systems for publishers. For instance, the system developers can implement BM25 as the similarity algorithm for the medicine subject domain while constructing the corpus using relatively short articles.

5.6 Summary: addressing research questions

The current research contributed to its achievements based on four major research questions stated in section 1.6. On the one hand, setting feasible and clear research questions supported this study to keep the research methodology on the correct track without deviating substantially from the important goals. On the other hand, understanding how these research questions are addressed by the study is important to have a rational idea of completeness of the proposed work. Therefore, the current section describes the way this research found answers for the research questions.

The first research question: "How does the importance of journal selection factors of OA journals vary in medicine and social sciences subject domains?" has begun to be addressed by the research in sections 3.2 and 3.3. Basically, a web-based survey was used to collect required information from authors to approach the question. Section 4.1.1 represents relevant results to the question, while section 5.1 describes the conclusions obtained. Accordingly, the peer-review status of a journal is identified as the most important factor for authors in both subject domains while the least importance was attributed to the number of annual subscribers. The mean importance determined for each factor has been given separately for the two domains, while revealing the existence of four factors which achieved significantly different importances in medicine compared to the social sciences. Finding strong correlations between factors separately in the two subject domains is also important for studying the different behaviour of publication factors across distinct subject domains.

Methodology of the second research question: "What is the most effective algorithm

for each subject domain, which suggests the most appropriate journal for input article abstract?" is described in section 3.4. The obtained results for addressing the research question is elaborated in section 4.2, while corresponding conclusions are explained in section 5.2. Results of average performance show that BM25 outperforms all the other algorithms against the test documents. By contrast, the unigram language measure demonstrates the least performance. Moreover, both of these observations are common in medicine and the social sciences subject domains. Findings of performance of five algorithms against sub-disciplines in the two subject domains also demonstrates a similar behavior, because BM25 works significantly better than other algorithms for almost all sub-disciplines across both subject domains. However, cosine similarity measure demonstrates better performance than other algorithms for sub-disciplines with higher density of technical vocabulary. The least performance against sub-disciplines is shown by the unigram language model. Performance of the algorithms is tested against average document lengths. According to the results, cosine similarity and SVM proves that they could increase performance when using shorter documents in medicine subject domain. The relationship of performance against the number of corpus journals demonstrates a moderate, positive, and statistically significant correlation for cosine measure and MNB in the social sciences domain, indicating their performance improvement with expanding the number of journals in training corpus.

The third research question: "Does the knowledge-based recommender component provide significant improvement of performance than the content-based recommender component alone?" is addressed by the sections 3.5, 3.6, and 3.7, particularly regarding the research methodology aspect. Section 4.4.2 reports relevant results for answering the third research question. Section 5.4 represents conclusions of the results obtained for comparing the C&K hybrid recommender system with the content-based component alone. Accordingly, the C&K recommender system demonstrates slight performance improvement over the content-based component alone. However, the difference in average performance for two recommender systems is not statistically significant across both subject domains. A number of potential reasons for this are

discussed by the dissertation. Moreover, section 5.7 suggests possible approaches of improving the performance of C&K recommender system over its present form.

The fourth research question: "Does the new journal recommender system provide appropriate suggestions compared to gold standards and authors?" has followed the methodology described in sections 3.5, 3.6, and 3.7 to obtain necessary results (see section 4.4.1) to address the question. Conclusions made for answering the research question are given by section 5.4. Accordingly, C&K recommender system suggests approximately 66.2% and 58.8% appropriate suggestions in medicine and social sciences domains respectively. Moreover, 35.5% medicine authors and 40.4% of social sciences authors indicate that the C&K recommender system suggested more appropriate journal(s) for their articles than the journal they actually published in. Compared to two gold standards defined in section 4.4.1, the average performance of C&K recommender system demonstrates 15% and 22.4% differences with gold standard 1 and gold standard 2 respectively in medicine subject domain. These percentages are considered as 18% and 28.4% for gold standard 1 and gold standard 2 respectively in the social sciences.

5.7 Future work

The current study suggests to extend the findings of the first author survey and the second author survey in two potential ways. On the one hand, it is desirable to find the precise reasons for explaining the obtained results. Reasons for why authors pay highest attention to the peer-review factor, and least attention to the reliability and demand category with respect to the other two categories could be revealed by another author survey. On the other hand, besides exploring the existence of significant differences between the gap between authors' publication needs and achieved factor standards, the current study does not explore what exactly cause these differences. This space could be bridged by examining the proposals of authors, stakeholders, editors, and publishers.

Another possible extension is to find exact reasons for explaining the behavior of the five algorithms in different contexts. For instance, one can further analyze the retrieved results given by the cosine measure for subject domains with higher technical vocabulary and subject domains with general vocabulary.

Increasing the performance of content-based recommender component can be investigated by the following numerous techniques. One possible option is optimizing two parameters of the BM25 algorithm. The recommender component currently uses two successful parameter values based on empirical evidence (Hong and Kim, 2016; Xu et al., 2016). However, this does not imply that the parameters cannot be tuned further to enhance accuracy of predictions. Nature of test and training documents or other insignificant characteristics in them at present could influence to change the performance of predictions parallel to adjusting the parameters. Therefore, a future study which investigates the variation of performance of prediction accuracy against the variation of two smoothing parameters could lead the way to enhance the current content-based recommender component. In addition to optimizing the parameters of current algorithm implemented in content-based component, a future research can also focus on replacing the current algorithm by a more effective one. On the one hand, the current study tested the suitability of only five algorithms, but one can think of more algorithms that may be appropriate for the system. On the other hand, adjusting the parameters of the other four algorithms which were tested could also lead to increased prediction accuracy over the prediction performance found at present.

Customizing factor weights based on authors' own opinion could be suggested as a method to examine whether the effect of knowledge-based component could lead to enhance the performance of C&K recommender system. Moreover, populating two corpora with more number of journals that represent the diverse range of subdisciplines belonging to medicine and social sciences can be considered as another possible approach for improving C&K recommender system.

Presence of vague terms in an input article abstract and corpus documents could

lead to a reduction in prediction accuracy of the content-based recommender component. A number of studies have experimentally proved the usefulness and influence of controlled vocabularies for numerous classification problems (Golub, 2006; Wijewickrema, 2014). A solution like introducing controlled vocabulary terms could avoid usage of vague terms in indexing and searching of the text. To illustrate, the term 'bank' has two distinct meanings under two different contexts. On the one hand, the term gives the idea of a financial organization. On the other hand, it means sloping raised land along the sides of a river. Therefore, there exists some possibility of misclassifying an input article as appropriate for publication in a geography journal, though it truly belongs to the sub-discipline – finance. C&K recommender system eradicates this problem to some extent as it uses separate corpus for distinct subject domains. Nevertheless, incidence of vague terms may also create trouble within sub-disciplines of a major subject domain. For instance, an article about blood cancer could be misclassified as an appropriate article for publication in a journal about oncology, while ignoring a number of journals belong to hematology. However, controlling 'blood cancer' as 'leukemia' in the sub-discipline - hematology could minimize this issue. Thus, associating a proper controlled vocabulary for each corpus representing distinct subject domains could enhance the accuracy of suggestions. This study proposes to use TemaTres[1], ThManager[2], Unilexicon[3], and Protégé[4] as tools for creating and managing controlled vocabularies to integrate with the C&K recommender system.

Dynamic nature of the factor values utilized by the knowledge-based recommender component to compare with authorss publication needs requires periodical updating of journal metadata. Even if, this nature does not heavily influence relatively matured, well established journals, the impact could make substantial changes in factors' values of novel journals. A future extension of the current system can consider the following proposal to overcome this problem.

[1] https://www.vocabularyserver.com/
[2] http://thmanager.sourceforge.net/
[3] https://unilexicon.com/
[4] https://protege.stanford.edu/

Usually, DOAJ supports for OAI - PMH (Open Archives Initiative Protocol for Metadata Harvesting). It arranges all information received from publishers according to an OAI compatible mode and acts as a data provider (DOAJ, 2019). Therefore, registering the proposed recommender system as an OAI service provider or retrieving data from an already registered service provider, will lead to the maintenance of up-to-date metadata set of OA journals for the purpose of C&K recommender system. Figures 5.7 and 5.8 illustrate the two possible ways of receiving metadata. Inabil-

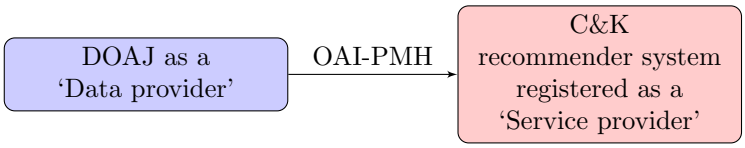


Figure 5.7: Metadata directly from DOAJ

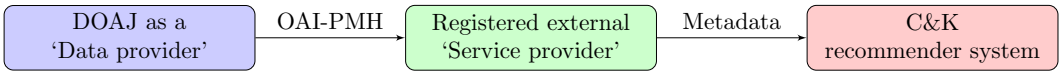


Figure 5.8: Metadata via external service provider

ity of searching across multiple databases is one of the main limitations associated with the C&K recommender system and other existing recommendation systems like JANE, eTBLAST, Elsevier Journal Finder and so on. Therefore, this study proposes to extend the C&K recommender system to allow to search and retrieve the most fitting journals from a wide range of databases. However, since different databases may use distinct journal metadata schemes, it would be necessary to develop a method to obtain uniform metadata records for all databases. The licensing issues between commercial publishers could also appear as an obstacle when attempting to connect with high-end commercial databases.

Finally, integrating user friendly interfaces can also be considered as an important aspect of a recommender system as the user's input interface gives the first impres-

sion of capabilities available. Designing an easy to understand, but inclusive of all functionalities system may require a number of cycles of improvements and user tests to ensure the interface suits well for its users. The current C&K recommender system uses Java console to feed test article abstracts and journals' bibliometric criteria to the system. This method requires a number of hits to execute the programme and inexperienced users may meet difficulties while navigating the way the system is functioning. However, this can easily be avoided by using a well designed interface. System output with suggested journals is also printed on the Java console window at present. This can be replaced by a compatible window, which prints suggested journals with hyperlinks to their official websites. This makes users comfortable since the strategy enables them to acquire more information of suggested journals.

This study has designed an appropriate prototype input interface (see figure 5.9) to feed needed information to the system, while designing another prototype output interface (see figure 5.10) to retrieve necessary information from the system.

All in all, the outcome of this dissertation is useful not only as an information system tool, but also it has contributed a number of significant findings to the field of scholarly publishing. Finding appropriate journals for submitting articles is challenging for some authors due to the incompetencies in the assistance they receive. The proposed recommender system was targeted at conquering these limitations and has demonstrated positive indications of success with respect to the authors and ideal ranking systems. Integrating potential improvements discussed under future works will further enhance the effectiveness of the system. The benefits of the new system will be acquired by some other stakeholders as discussed in the beginning of the dissertation. Therefore, the contribution of this dissertation is expected to make a noteworthy advance in scholarly publishing in the time to come.

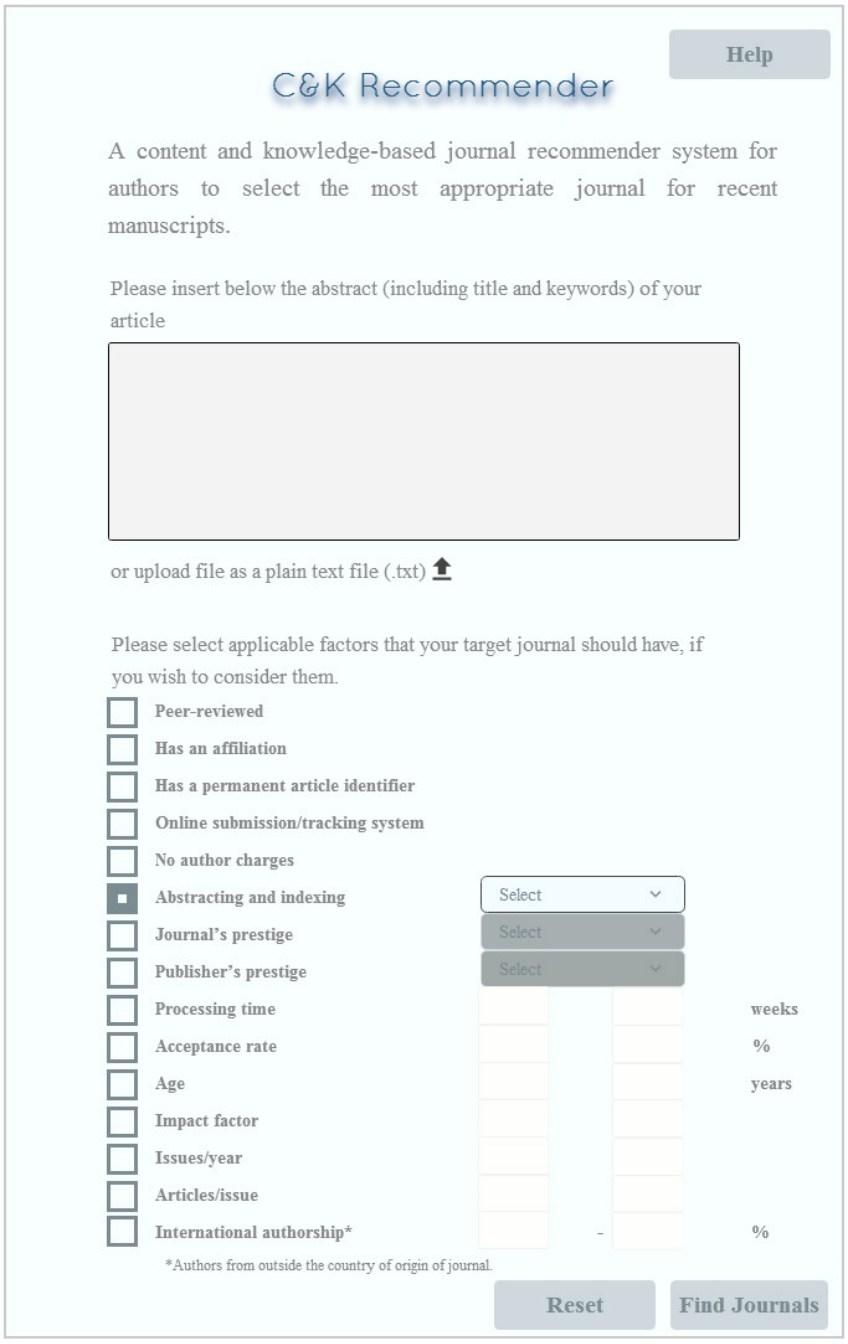


Figure 5.9: User's input interface

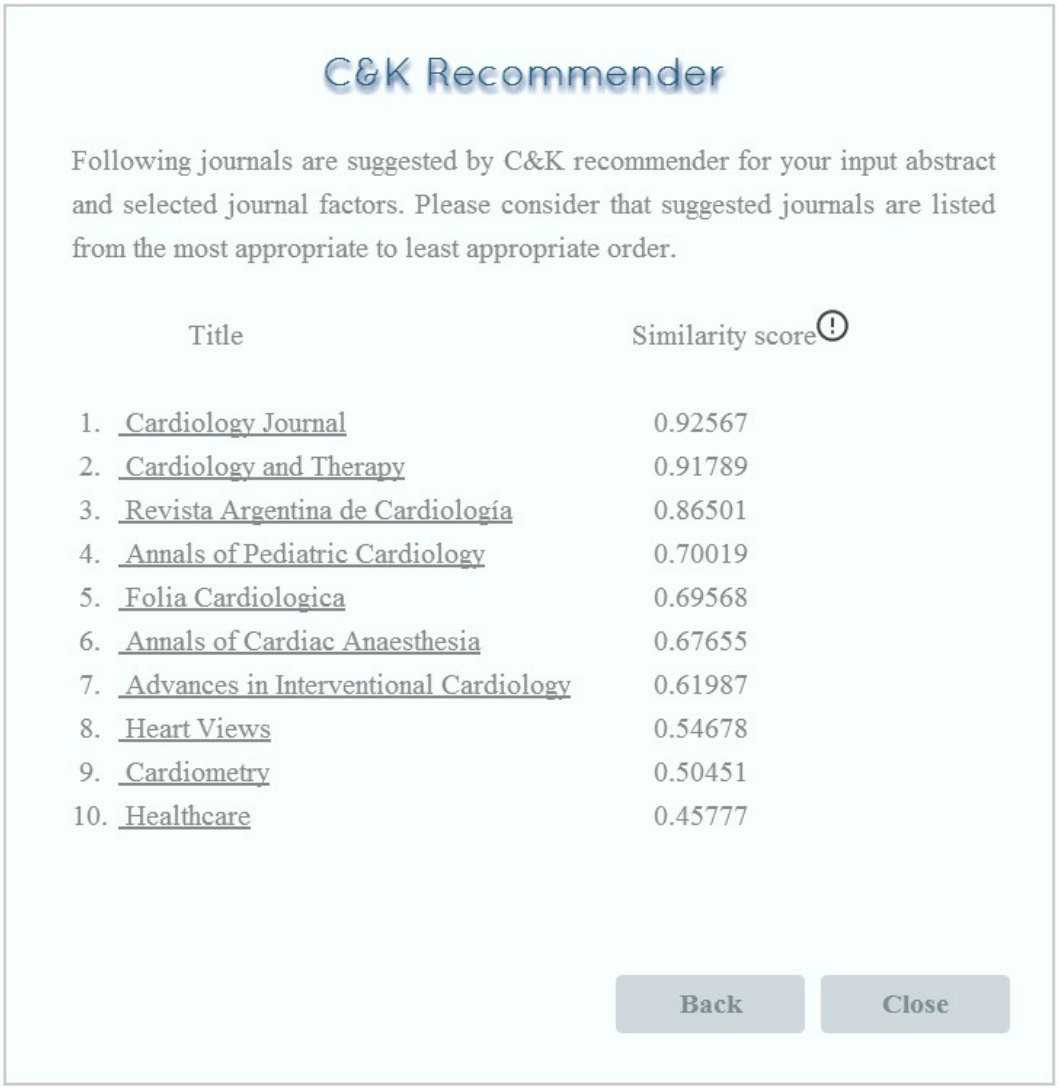


Figure 5.10: User's output interface

Bibliography

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

Alhoori, H. and Furuta, R. (2017). Recommendation of scholarly venues based on dynamic user interests. *Journal of Informetrics*, 11(2):553–563.

Amin, M. and Mabe, M. (2004). Impact factors: use and abuse. *International Journal of Environmental Science and Technology:(IJEST)*, 1(1):1.

Apache Software Foundation (2010). Similarity (Lucene 3.0.3 API). https://lucene.apache.org/core/3_0_3/api/core/org/apache/lucene/search/Similarity.html. [accessed May 2017].

Apache Software Foundation (2017). BM25 Similarity (Lucene 6.5.0 API). https://lucene.apache.org/core/6_5_0/core/org/apache/lucene/search/similarities/BM25Similarity.html. [accessed May 2017].

Arendell, T. and Reinharz, S. (1995). Feminist methods in social research. *Social Forces*, 73(4):1636.

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Banko, M. and Brill, E. (2001). Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. In *Proceedings of the First International Conference on Human Language*

Technology Research, HLT '01, pages 1–5, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, 3(2):77–85.

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, 68(5):314–316.

Białecki, A., Muir, R., and Ingersoll, G. (2012). Apache Lucene 4. In *SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 17–24.

Björk, B.-C. and Holmström, J. (2006). Benchmarking scientific journals from the submitting author's viewpoint. *Learned Publishing*, 19(2):147–155.

Björk, B.-C. and Öörni, A. (2009). A method for comparing scholarly journals as service providers to authors. *Serials Review*, 35(2):62–69.

Bogers, T. and van den Bosch, A. (2007). Comparing and evaluating information retrieval algorithms for news recommendation. In *Proceedings of the 2007 ACM Conference on Recommender Systems - RecSys '07*. ACM Press.

Boukhris, I. and Ayachi, R. (2014). A novel personalized academic venue hybrid recommender. In *2014 IEEE 15th International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 465–470.

Bradshaw, C. J. A. and Brook, B. W. (2016). How to rank journals. *PLOS ONE*, 11(3):1–15.

Broome, M. E. (2007). A rose by any other name is still a rose: Assessing journal quality. *Nursing Outlook*, 55(4):163–164.

Bröchner, J. and Björk, B.-C. (2008). Where to submit? Journal choice by construction management authors. *Construction Management and Economics*, 26(7):739–749.

Burke, R. (2007). Hybrid web recommender systems. In Peter Brusilovsky, Alfred Kobsa, W. N., editor, *The Adaptive Web*, Lecture Notes in Computer Science, pages 377–408. Springer Berlin Heidelberg.

Busa-Fekete, R., Szarvas, G., Élteto, T., and Kégl, B. (2012). An apple-to-apple comparison of Learning-to-rank algorithms in terms of Normalized Discounted Cumulative Gain. In Raedt, D., L., Bessiere, C., Dubois, D., Doherty, P., Frasconi, P., Heintz, F., Lucas, and P., editors, *ECAI 2012 - 20th European Conference on Artificial Intelligence: Preference Learning: Problems and Applications in AI Workshop*, volume 242, Montpellier, France. Ios Press.

Busa-Fekete, R., Szörényi, B., Dembczynski, K., and Hüllermeier, E. (2015). Online F-Measure optimization. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 595–603. Curran Associates, Inc.

Carterette, B. and Jones, R. (2008). Evaluating search engines by modeling the relationship between relevance and clicks. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 217–224. Curran Associates, Inc.

Chapelle, O., Metlzer, D., Zhang, Y., and Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 621–630, New York, NY, USA. ACM.

Chazdon, R. L., Finegan, B., Capers, R. S., Salgado-Negret, B., Casanoves, F., Boukili, V., and Norden, N. (2009). Composition and dynamics of functional groups of trees during tropical forest succession in Northeastern Costa Rica. *Biotropica*, 42(1):31–40.

Cheung, C.-K. (2008). Audience matters: A study of how authors select educational journals. *The Asia-Pacific Education Researcher*, 17(2):191–201.

Chu, H. and Rosenthal, M. (1996). Search engines for the World Wide Web: A comparative study and evaluation methodology. In *Proceedings of the Annual Meeting-American Society for Information Science*, volume 33, pages 127–135.

Clarke, C. L. A., Cormack, G. V., Laszlo, M., Lynam, T. R., and Terra, E. L. (2002). The impact of corpus size on question answering performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 369–370, New York, NY, USA. ACM.

Clarke, S. J. and Willett, P. (1997). Estimating the recall performance of Web search engines. *Aslib Proceedings*, 49(7):184–189.

Cope, B. and Phillips, A., editors (2014). *The Future of the Academic Journal*. Chandos Publishing Series. Elsevier Science.

Corrêa, R. X., Abdelnoor, R. V., Faleiro, F. G., Cruz, C. D., Moreira, M. A., and Barros, E. G. D. (1999). Genetic distances in soybean based on RAPD markers. *Bragantia*, 58(1):15–22.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Dalton, M. (2013). A dissemination divide? The factors that influence the journal selection decision of Library & Information Studies (LIS) researchers and practitioners. *Library and Information Research*, 37(115):33–57.

Ding, W. and Marchionini, G. (1996). A comparative study of web search service performance. In *Proceedings of the ASIST Annual Meeting*, volume 33, pages 136–42.

Dixon-Woods, M. and Tarrant, C. (2009). Why do people cooperate with medical research? Findings from three studies. *Social Science & Medicine*, 68(12):2215–2222.

DOAJ (2019). OAI-PMH. https://doaj.org/features. Accessed: 2019-01-17.

Dorta-González, P. and Dorta-González, M. I. (2013). Comparing journals from different fields of science and social science through a JCR subject categories normalized impact factor. *Scientometrics*, 95(2):645–672.

Eekhout, I., de Boer, R. M., Twisk, J. W. R., de Vet, H. C. W., and Heymans, M. W. (2012). Missing data. *Epidemiology*, 23(5):729–732.

Elsevier (2017). Scopus content coverage guide. https://www.elsevier.com/__data/assets/pdf_file/0007/69451/0597-Scopus-Content-Coverage-Guide-US-LETTER-v4-HI-singles-no-ticks.pdf. Accessed: 2017-04-19.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Fangerau, H. (2004). Finding European bioethical literature: an evaluation of the leading abstracting and indexing services. *Journal of Medical Ethics*, 30(3):299–303.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Fisher, R. A. (1934). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd, 5 edition.

Fleming, C. M. and Bowden, M. (2009). Web-based surveys as an alternative to traditional mail methods. *Journal of Environmental Management*, 90(1):284–292.

Forrester, A., Björk, B.-C., and Tenopir, C. (2017). New web services that help authors choose journals. *Learned Publishing*, 30(4):281–287.

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., and Trigg, L. (2009). Weka–a machine learning workbench for data mining. In *Data Mining and Knowledge Discovery Handbook*, pages 1269–1277. Springer.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479.

Gatta, R., Vallati, M., Bari, B. D., and Ozsahin, M. (2014). The impact of different training sets on medical documents classification. In *Proceedings of the 3rd International Conference on Artificial Intelligence and Assistive Medicine - Volume 1213*, AIAM'14, pages 1–5, Aachen, Germany, Germany. CEUR-WS.org.

Gennaro, C., Amato, G., Bolettieri, P., and Savino, P. (2010). An approach to content-based image retrieval based on the Lucene search engine library. In *Research and Advanced Technology for Digital Libraries*, pages 55–66. Springer Berlin Heidelberg.

Golub, K. (2006). Automated subject classification of textual Web pages, based on a controlled vocabulary: Challenges and recommendations. *New Review of Hypermedia and Multimedia*, 12(1):11–27.

González-Betancor, S. M. and Dorta-González, P. (2017). An indicator of the impact of journals based on the percentage of their highly cited publications. *Online Information Review*, 41(3):398–411.

González-Pereira, B., Guerrero-Bote, V. P., and Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of informetrics*, 4(3):379–391.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871.

Gräßer, F., Malberg, H., Zaunseder, S., Beckert, S., Köster, D., Schmitt, J., Klinik, S. A., and für Dermatologie, P. (2016). Application of recommender system methods for therapy decision support. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6.

Groneberg, D. A. (2018). Social sciences research in the central european city of wrocław: A density-equalizing mapping analysis. *PLOS ONE*, 13(10):e0205094.

Gutknecht, C. (2014). *Where to publish? Development of a recommender system*

for academic publishing. Master thesis, University of Applied Sciences and Arts Northwestern Switzerland.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations Newsletter*, 11(1):10–18.

Hart, A. (2001). Mann-Whitney test is not just a test of medians: differences in spread can be important. *BMJ*, 323(7309):391–393.

He, B. and Ounis, I. (2003). A study of parameter tuning for term frequency normalization. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 10–16, New York, NY, USA. ACM.

He, C. and Pao, M. L. (1986). A discipline-specific journal selection algorithm. *Information Processing & Management*, 22(5):405–416.

Hearst, M. A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M. A., and Ye, J. (2007). BioText Search Engine: Beyond abstract search. *Bioinformatics*, 23(16):2196–2197.

Heintzelman, M. and Nocetti, D. (2009). Where should we submit our manuscript? An analysis of journal submission strategies. *Berkeley Electronic Journal of Economic Analysis and Policy (Advances)*, 9(1).

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.

Hong, B. and Kim, Y. (2016). A weighted question retrieval model using descriptive information in community question answering. In *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*, RACS '16, pages 35–39, New York, NY, USA. ACM.

Islam, A. and Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, 2(2):1–25.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Johnson, A. (2008). How MoreLikeThis works in Lucene. http://cephas.net/blog/2008/03/30/how-morelikethis-works-in-lucene/.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information Processing & Management*, 36(6):809–840.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Kagie, M., van Wezel, M., and Groenen, P. J. (2008). A graphical shopping interface based on product attributes. *Decision Support Systems*, 46(1):265–276.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1):31–36.

Kaiser, H. F. and Rice, J. (1974). Little Jiffy, Mark Iv. *Educational and Psychological Measurement*, 34(1):111–117.

Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., and Robertson, S. (2007). INEX 2007 evaluation measures. In *Focused Access to XML Documents*, pages 24–33. Springer Berlin Heidelberg.

Kang, N., Doornenbal, M. A., and Schijvenaars, R. J. (2015). Elsevier journal finder: Recommending journals for your paper. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pages 261–264, New York, NY, USA. ACM.

Kibriya, A. M., Frank, E., Pfahringer, B., and Holmes, G. (2004). Multinomial naive Bayes for text categorization revisited. In *Lecture Notes in Computer Science*, pages 488–499. Springer Berlin Heidelberg.

Kim, H.-Y. (2017). Statistical notes for clinical researchers: chi-squared test and Fisher's exact test. *Restorative Dentistry & Endodontics*, 42(2):152–155.

Klein, D. B. and Chiang, E. (2004). The Social Science Citation Index: A black box-with an ideological bias? *Econ Journal Watch*, 1(1):134–165.

Klingner, J. K., Scanlon, D., and Pressley, M. (2005). How to publish in scholarly journals. *Educational Researcher*, 34(8):14–20.

Knight, L. V. and Steinbach, T. A. (2008). Selecting an appropriate publication outlet: A comprehensive model of journal selection criteria for researchers in a broad range of academic disciplines. *International Journal of Doctoral Studies*, 3:59–79.

Kosaki, K., Jones, M. C., and Stayboldt, C. (1996). Zimmer phocomelia: delineation by principal coordinate analysis. *American Journal of Medical Genetics*, 66(1):55–59.

Krause, E. F. (1973). Taxicab geometry. *The Mathematics Teacher*, 66(8):695–706.

Kumar, S. and Gupta, P. (2015). Comparative analysis of intersection algorithms on queries using precision, recall and F-score. *International Journal of Computer Applications*, 130(7):28–36.

Lagrange, J. L. (1853). *Mécanique analytique*, volume 1. Mallet-Bachelier.

Landry, R., Amara, N., and Lamari, M. (2001). Utilization of social science research knowledge in canada. *Research Policy*, 30(2):333–349.

Langston, L. (1996). Scholarly communication and electronic publication: Implications for research, advancement, and promotion. In *Untangling the Web: Proceedings of the Conference Sponsored by the Librarians Association of the University of California, Santa Barbara and Friends of the UCSB Library*, pages 1–10.

Larsen, P. O. and von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3):575–603.

Lewallen, L. P. and Crane, P. B. (2010). Choosing a publication venue. *Journal of Professional Nursing*, 26(4):250–254.

Liang, Y., Chang, D., Huang, Y., Hu, S., Song, R., and Sun, D. (2013). Exploration and fulfillment of search engine in economic model resource platform subsystem based on Lucene search engine. In *LISS 2013*, pages 1017–1022. Springer Berlin Heidelberg.

Lorenzo-Seva, U. and Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, 38(1):88–91.

Lu, Z., Xie, N., and Wilbur, W. J. (2009). Identifying related journals through log analysis. *Bioinformatics*, 25(22):3038–3039.

Luong, H., Huynh, T., Gauch, S., Do, L., and Hoang, K. (2012). Publication venue recommendation using author network's publication history. In *Intelligent Information and Database Systems*, pages 426–435. Springer Berlin Heidelberg.

Mansfield, S. and Mcardle, B. H. (1998). Dietary composition of Gambusia affinis (Family Poeciliidae) populations in the northern Waikato region of New Zealand. *New Zealand Journal of Marine and Freshwater Research*, 32(3):375–383.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., and Yarkoni, T. (2016). Point of view: How open science helps researchers succeed. *eLife*, 5:1–19.

Mudrak, B. (2015). JournalGuide: bringing authors and journals together. *Learned Publishing*, 28(2):147–149.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71.

Nanda, M. A., Seminar, K. B., Nandika, D., and Maddu, A. (2018). A comparison study of kernel functions in the support vector machine and its application for termite detection. *Information*, 9(5):1–14.

Nascimento, A., Smith, D., Pereira, S., de Araújo, M. S., Silva, M., and Mariani, A. (2000). Integration of varying responses of different organisms to water and sediment quality at sites impacted and not impacted by the petroleum industry. *Aquatic Ecosystem Health & Management*, 3(4):449–458.

Nelson, A. (2009). Inclusion: the politics of difference in medical research. *Social Identities*, 15(5):741–743.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & Evaluation in Higher Education*, 33(3):301–314.

Ogurtsov, M., Nagovitsyn, Y., Kocharov, G., and Jungner, H. (2002). Long-period cycles of the sun's activity recorded in direct solar data and proxies. *Solar Physics*, 211(1):371–394.

Olden, J. D., Poff, N. L., and Bestgen, K. R. (2006). Life-history strategies predict fish invasions and extirpations in the Colorado river basin. *Ecological Monographs*, 76(1):25–40.

Ortiz, D., Myers, D., Walls, E., and Diaz, M.-E. (2005). Where do we stand with newspaper data? *Mobilization: An International Quarterly*, 10(3):397–419.

Oster, S. (1980). The optimal order for submitting manuscripts. *The American Economic Review*, 70(3):444–448.

Özçakar, L., Franchignoni, F., Kara, M., and Muñoz, S. L. (2012). Choosing a scholarly journal during manuscript submission: the way how it rings true for physiatrists. *European Journal of Physical and Rehabilitation Medicine*, 48(4):643–647.

Pavoine, S., Vallet, J., Dufour, A.-B., Gachet, S., and Daniel, H. (2009). On the challenge of treating various types of variables: Application for improving the measurement of functional diversity. *Oikos*, 118(3):391–402.

Pearson, K. and Pearson, E. S. (1922). On polychoric coefficients of correlation. *Biometrika*, 14(1-2):127–156.

Pham, M. C., Cao, Y., Klamma, R., and Jarke, M. (2011). A clustering approach for collaborative filtering recommendation using social network analysis. *Journal of Universal Computer Science*, 17(4):583–604.

Pillar, V. D. and Sosinski, E. E. (2003). An improved method for searching plant functional types by numerical analysis. *Journal of Vegetation Science*, 14(3):323–332.

Pirro, G. and Talia, D. (2007). An approach to ontology mapping based on the Lucene search engine library. In *18th International Workshop on Database and Expert Systems Applications (DEXA 2007)*, pages 407–411.

Platt, J. (1999a). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.

Platt, J. C. (1999b). Fast training of support vector machines using sequential minimal optimization. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods*, pages 185–208. MIT Press, Cambridge, MA, USA.

Poff, N. L., Olden, J. D., Vieira, N. K., Finn, D. S., Simmons, M. P., and Kondratieff, B. C. (2006). Functional trait niches of North American lotic insects: Traits-based ecological applications in light of phylogenetic relationships. *Journal of the North American Benthological Society*, 25(4):730–755.

Pong, J. Y.-H., Kwok, R. C.-W., Lau, R. Y.-K., Hao, J.-X., and Wong, P. C.-C.

(2007). A comparative study of two automatic document classification methods in a library setting. *Journal of Information Science*, 34(2):213–230.

Prabowo, R., Jackson, M., Burden, P., and Knoell, H. . (2002). Ontology-based automatic classification for Web pages: Design, implementation and evaluation. In *Proceedings of the Third International Conference on Web Information Systems Engineering, 2002. WISE 2002.*, pages 182–191.

Priebe, M. M., Sherwood, A. M., Thornby, J. I., Kharas, N. F., and Markowski, J. (1996). Clinical assessment of spasticity in spinal cord injury: A multidimensional problem. *Archives of Physical Medicine and Rehabilitation*, 77(7):713–716.

Regazzi, J. J. and Aytac, S. (2008). Author perceptions of journal quality. *Learned Publishing*, 21(3):225–235.

Ricci, F., Rokach, L., and Shapira, B. (2015). Recommender systems: Introduction and challenges. In *Recommender Systems Handbook*, pages 1–34. Springer US.

Rison, R. A., Shepphird, J. K., and Kidd, M. R. (2017). How to choose the best journal for your case report. *Journal of Medical Case Reports*, 11(1):198.

Rollins, J., McCusker, M., Carlson, J., and Stroll, J. (2017). Manuscript matcher: A content and bibliometrics-based scholarly journal recommendation system. In *Proceedings of the Fifth Workshop on Bibliometric-enhanced Information Retrieval (BIR) co-located with the 39th European Conference on Information Retrieval (ECIR 2017), Aberdeen, UK, April 9th, 2017.*, pages 18–29.

Ronte, H. (2001). The impact of technology on publishing. *Publishing Research Quarterly*, 16(4):11–22.

Rousseau, S. and Rousseau, R. (2012). Interactions between journal attributes and authors' willingness to wait for editorial decisions. *Journal of the American Society for Information Science and Technology*, 63(6):1213–1225.

Rowlands, I. and Nicholas, D. (2006). The changing scholarly communication landscape: An international survey of senior researchers. *Learned Publishing*, 19(1):31–55.

Rowlands, I., Nicholas, D., and Huntington, P. (2004). Scholarly communication in the digital environment: What do authors want? *Learned Publishing*, 17(4):261–273.

Rubin, D. B., editor (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc.

Sanchez Bocanegra, C. L., Sevillano Ramos, J. L., Rizo, C., Civit, A., and Fernandez-Luque, L. (2017). HealthRecSys: A semantic content-based recommender system to complement health videos. *BMC Medical Informatics and Decision Making*, 17(1):63.

Schuemie, M. J. and Kors, J. A. (2008). Jane: Suggesting journals, finding experts. *Bioinformatics*, 24(5):727–728.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to Information Retrieval*, volume 39. Cambridge University Press.

Scully, C. and Lodge, H. (2005). Impact factors and their significance; overrated or misused? *British Dental Journal*, 198(7):391–393.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, 314(7079):497–497.

Seife, C. (2014). *Virtual Unreality: Just Because the Internet Told You, how Do You Know It's True?* Penguin Publishing Group.

Shafi, S. and Rather, R. (2005). Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. *Webology*, 2(2).

Sharman, A. (2015). Where to publish. *The Annals of The Royal College of Surgeons of England*, 97(5):329–332.

Shavers-Hornaday, V., Lynch, C., Burmeister, L., and Torner, J. (1997). Why are African Americans under-represented in medical research studies? Impediments to participation. *Ethnicity & Health*, 2(1-2):31–45.

Shokraneh, F., Ilghami, R., Masoomi, R., and Amanollahi, A. (2012). How to select a journal to submit and publish your biomedical paper? *BioImpacts: BI*, 2(1):61–68.

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. (1999). Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 33(1):6–12.

Sixto, J., Almeida, A., and López-de Ipiña, D. (2016). Improving the sentiment analysis process of Spanish tweets with BM25. In *International Conference on Applications of Natural Language to Information Systems*, pages 285–291. Springer.

Solomon, D. and Björk, B.-C. (2016). Article processing charges for open access publication—the situation for research intensive universities in the USA and Canada. *PeerJ*, 4:e2264.

Solomon, D. J. and Björk, B.-C. (2012). Publication fees in open access publishing: Sources of funding and factors influencing choice of journal. *Journal of the American Society for Information Science and Technology*, 63(1):98–107.

Solomon, D. J. and Björk, B.-C. (2012). A study of open access journals using article processing charges. *Journal of the American Society for Information Science and Technology*, 63(8):1485–1495.

Solomon, D. J., Laakso, M., and Björk, B.-C. (2013). A longitudinal comparison of citation rates and growth among open access journals. *Journal of Informetrics*, 7(3):642–650.

Søreide, K. and Winter, D. C. (2010). Global survey of factors influencing choice of surgical journal for manuscript submission. *Surgery*, 147(4):475–480.

Tallarida, R. J. and Murray, R. B. (1987). Mann-whitney test. In *Manual of Pharmacologic Calculations*, pages 149–153. Springer New York.

Tattersall, A., editor (2016). *Altmetrics: A practical guide for librarians, researchers and academics*. London: Facet Publishing.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., Robson, R., Thabane, M., Giangregorio, L., and Goldsmith, C. H. (2010). A tutorial on pilot studies: The what, why and how. *BMC Medical Research Methodology*, 10(1):1–10.

Thompson, S. K. (2012). *Sampling*. John Wiley & Sons, Inc.

Usmani, T. A., Pant, D., and Bhatt, A. K. (2012). A comparative study of google and bing search engines in context of precision and relative recall parameter. *International Journal on Computer Science and Engineering*, 4(1):21–34.

van Teijlingen, E. and Hundley, V. (2002). The importance of pilot studies. *Nursing Standard*, 16(40):33–36.

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing & Management*, 40(4):677–691.

Vitali, C., Moutsopoulos, H. M., and Bombardieri, S. (1994). The European Community Study Group on diagnostic criteria for Sjögren's syndrome. Sensitivity and specificity of tests for ocular and oral involvement in Sjögren's syndrome. *Annals of the Rheumatic Diseases*, 53(10):637–647.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2):365–391.

Welch, S. J. (2012). Selecting the right journal for your submission. *Journal of Thoracic Disease*, 4(3):336–338.

Wijewickrema, C. M. (2014). Impact of an ontology for automatic text classification. *Annals of Library and Information Studies*, 61(4):263–272.

Wijewickrema, M. and Petras, V. (2017). Journal selection criteria in an open access environment: A comparison between the medicine and social sciences. *Learned Publishing*, 30(4):289–300.

Wijewickrema, M., Petras, V., and Dias, N. (in press 2019). Selecting a text similarity measure for a content-based recommender system: A comparison in two corpora. *The Electronic Library*.

Wijewickrema, P. K. C. M. (2015). A three dimensional model for selecting the most appropriate journal outlet for manuscript submission. In *International Conference on Social Science Research, ICSSR*, pages 488–504.

Willett, P. (2006). The Porter stemming algorithm: Then and now. *Program*, 40(3):219–223.

Williams, B., Onsman, A., and Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Journal of Emergency Primary Health Care (JEPHC)*, 8(3):1–13.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, second edition.

Wood, A. M., White, I. R., and Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials: Journal of the Society for Clinical Trials*, 1(4):368–376.

Wren, J. D., Hicks, J. M., Errami, M., and Garner, H. R. (2007). eTBLAST: A web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic Acids Research*, 35(suppl_2):W12–W15.

Xu, B., Lin, H., Hao, H., Yang, Z., Wang, J., and Zhang, S. (2016). Generating user-oriented text summarization based on social networks using topic models. In Li, Y., Xiang, G., Lin, H., and Wang, M., editors, *Social Media Processing*, pages 186–193, Singapore. Springer Singapore.

Xu, S., McCusker, J., and Krauthammer, M. (2008). Yale Image Finder (YIF): A new search engine for retrieving biomedical images. *Bioinformatics*, 24(17):1968–1970.

Yang, Z. and Davison, B. D. (2012). Venue recommendation: Submitting your paper with style. In *2012 11th International Conference on Machine Learning and Applications*, volume 1 of *ICMLA '12*, pages 681–686.

Yong Wang, J. H. and Tang, B. (2003). Classification of Web documents using a naive Bayes method. In *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*, pages 560–564.

Zhai, C. (2008). Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1(1):1–141.

Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.

Zhou, D.-P. and Xie, K.-L. (2007). Lucene search engine. *Jisuanji Gongcheng/ Computer Engineering*, 33(18):95–96.

Zhou, W., Smalheiser, N. R., and Yu, C. (2006). A tutorial on information retrieval: basic terms and concepts. *Journal of Biomedical Discovery and Collaboration*, 1(1):2.

Ziobrowski, A. and Gibler, K. (2000). Factors academic real estate authors consider when choosing where to submit a manuscript for publication. *Journal of Real Estate Practice and Education*, 3(1):43–54.

Appendices

Appendix A

First author survey: Manuscript submission considerations

A.1 Email invitation for first author survey

Subject: Manuscript Submission

Dear Sir/Madam,

As a part of my PhD research at the Berlin School of Library and Information Science, Humboldt University of Berlin, Germany, I wish to develop a novel journal ranking metric to select an appropriate journal for submitting manuscripts.

Therefore, I kindly invite you to assess the importance of 16 aspects that might influence your decision to submit an article to a particular journal.

This questionnaire is web-based. Please use the following link to start the survey (length of survey: approximately 10 minutes):

https://umfrage.hu-berlin.de/index.php/529334?newtest=Y

Your responses will be kept in absolute confidence and will be used for academic purposes

only. If you require additional information or have any questions, then please feel free to contact me.

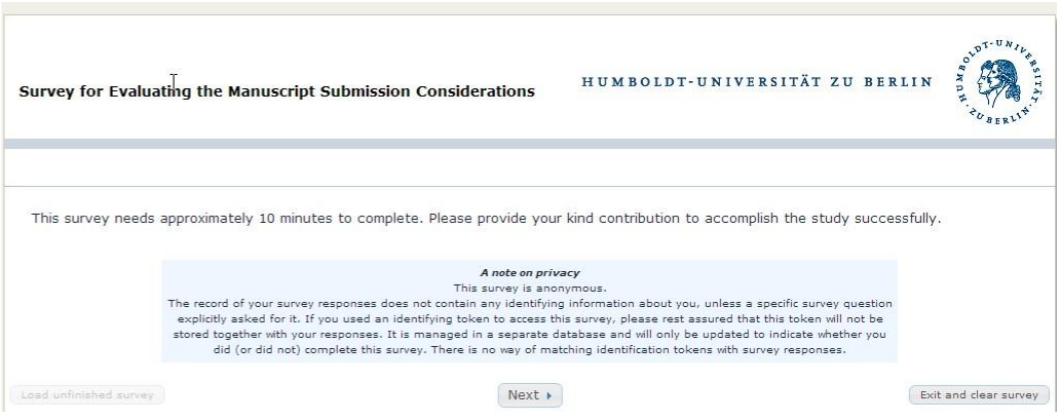
Thank you for taking time to assist me in my educational endeavors.

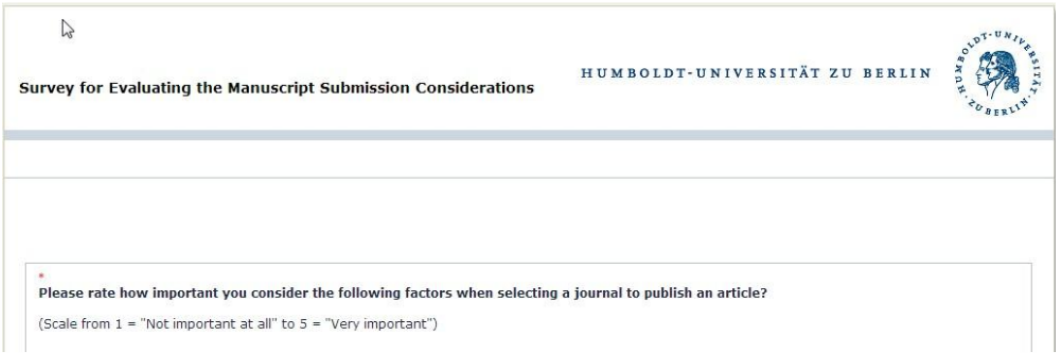
Sincerely,

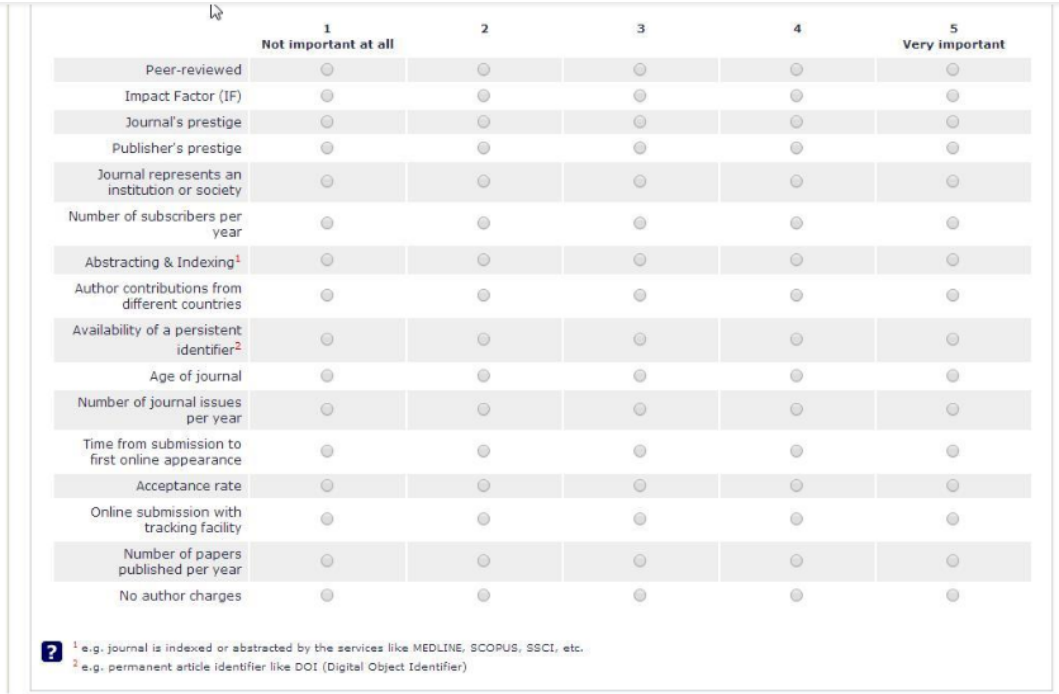
P. K. C. M. Wijewickrema

Email: wijewicm@student.hu-berlin.de

A.2 Questionnaire

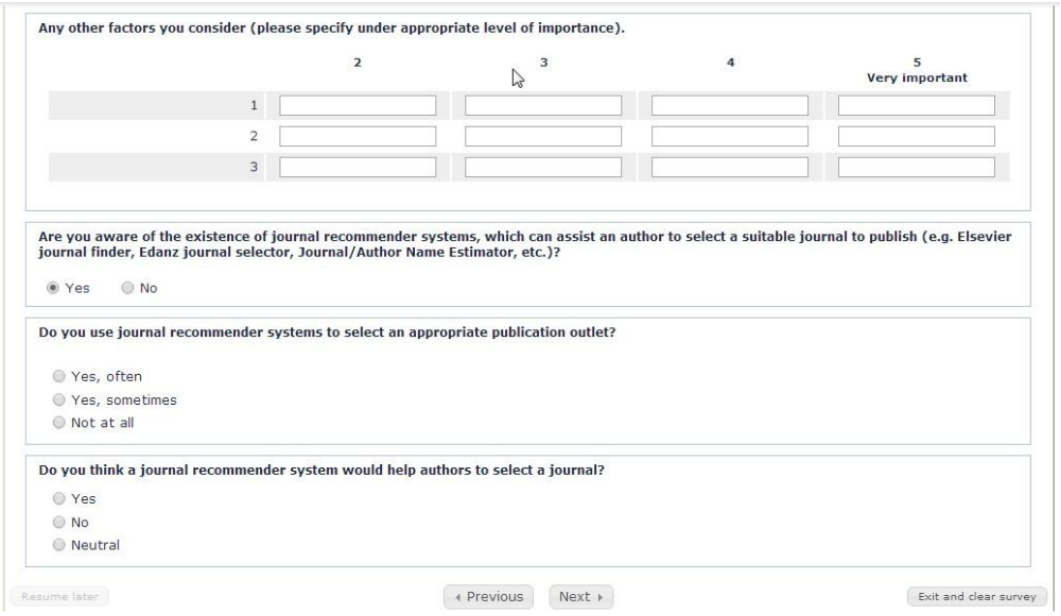






	1 Not important at all	2	3	4	5 Very important
Peer-reviewed					
Impact Factor (IF)					
Journal's prestige					
Publisher's prestige					
Journal represents an institution or society					
Number of subscribers per year					
Abstracting & Indexing[1]					
Author contributions from different countries					
Availability of a persistent identifier[2]					
Age of journal					
Number of journal issues per year					
Time from submission to first online appearance					
Acceptance rate					
Online submission with tracking facility					
Number of papers published per year					
No author charges					

[1] e.g. journal is indexed or abstracted by the services like MEDLINE, SCOPUS, SSCI, etc.
[2] e.g. permanent article identifier like DOI (Digital Object Identifier)



Any other factors you consider (please specify under appropriate level of importance).

	2	3	4	5 Very important
1				
2				
3				

Are you aware of the existence of journal recommender systems, which can assist an author to select a suitable journal to publish (e.g. Elsevier journal finder, Edanz journal selector, Journal/Author Name Estimator, etc.)?

(•) Yes ( ) No

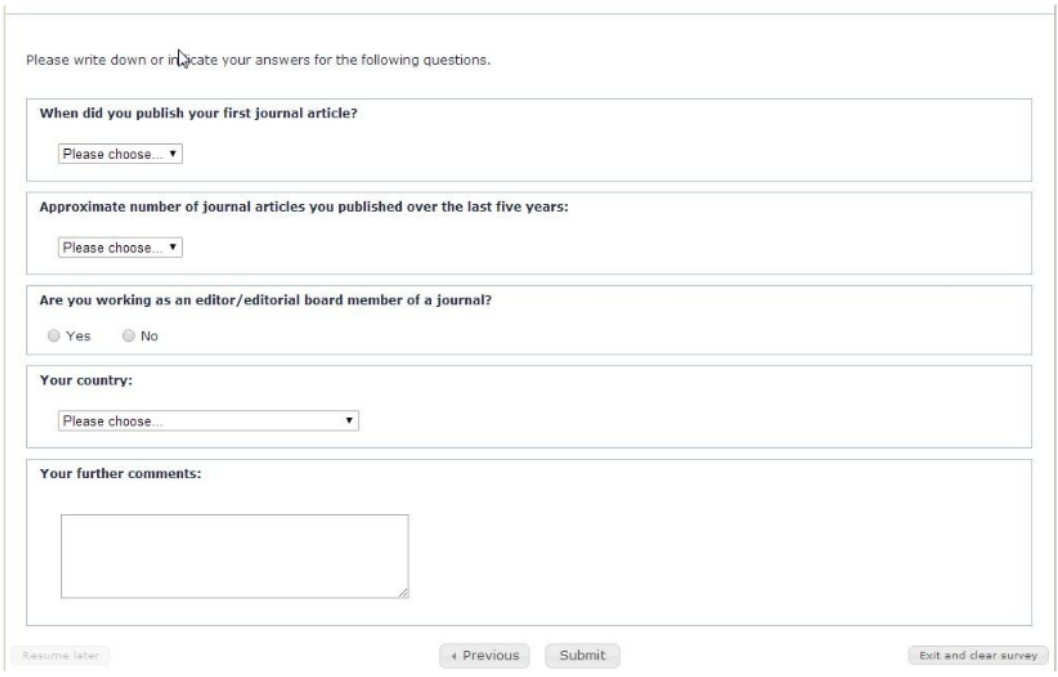
Do you use journal recommender systems to select an appropriate publication outlet?

( ) Yes, often
( ) Yes, sometimes
( ) Not at all

Do you think a journal recommender system would help authors to select a journal?

( ) Yes
( ) No
( ) Neutral

Resume later | ‹ Previous | Next › | Exit and clear survey



Please write down or indicate your answers for the following questions.

When did you publish your first journal article?

Please choose... ▾

Approximate number of journal articles you published over the last five years:

Please choose... ▾

Are you working as an editor/editorial board member of a journal?

( ) Yes ( ) No

Your country:

Please choose... ▾

Your further comments:

Resume later | ‹ Previous | Submit | Exit and clear survey

HUMBOLDT-UNIVERSITÄT ZU BERLIN

The survey is now complete. Thank you once again for your time!

Please be assured that your responses will remain confidential. If you need additional information or have any questions, please feel free to contact me on wijewicm@student.hu-berlin.de

Appendix B

Email invitation to editors

This email views the general format of an email, which was sent to editors for collecting journal metadata. However, the requested metadata information was different for some journals based on the availability of information in the primary sources.

Subject: Request to Provide Journal Information Dear Sir/Madam,

As a part of my PhD research at the Berlin School of Library and Information Science, Humboldt University of Berlin, Germany, I wish to develop a novel journal ranking metric based on 16 factors of a journal and ultimately to develop a recommender system, which can assist authors to select an appropriate open access journal outlet to submit their manuscripts.

The "*Name of the journal*" is included as a corpus journal in this novel system. To complete the required factors of the corpus journals, I kindly request you to provide following information for research articles published in the year 2016.

1. Acceptance rate (ratio between accepted and submitted articles) for general issues
2. Number of full-text article downloads

Your responses will be kept in absolute confidence and will be used for academic purposes only. If you require additional information or have any questions, then please feel free to contact me.

Thank you for taking time to assist me in my educational endeavors.

Sincerely,

P. K. C. M. Wijewickrema

Email: wijewicm@student.hu-berlin.de

Appendix C

Second author survey: Collecting data to configure the recommender system

C.1 Email invitation for second author survey

Subject: Journal Recommendation

Dear Sir/Madam,

As a part of my PhD research at the Berlin School of Library and Information Science, Humboldt University of Berlin, Germany, I developed a novel journal recommender system, which could assist authors to select an appropriate journal for submitting manuscripts. This survey aims to collect data to configure the recommender system, which given (a) a manuscript text and (b) journal characteristics recommends the most fitting journals for your text.

The recommender system is based on 15 journal characteristics. This survey asks about the journal characteristics, which contributed to your decision to select an appropriate journal to publish your article: "*Name of the article*". Please estimate whether and to what extent each characteristic contributed to your submission decision.

This questionnaire is web-based. Please use the following link to start the survey (length of survey: approximately 10 minutes):

https://umfrage.hu-berlin.de/index.php/survey/index/sid/936774/newtest/Y/lang/en

The next part of this survey would send you appropriate journals suggested by the recommender system based on your answers for this survey. This succeeding part allows you to rank the appropriateness of suggested journals for your article based on your own opinion.

Your responses will be kept in absolute confidence and will be used for academic purposes only. If you require additional information or have any questions, then please feel free to contact me.

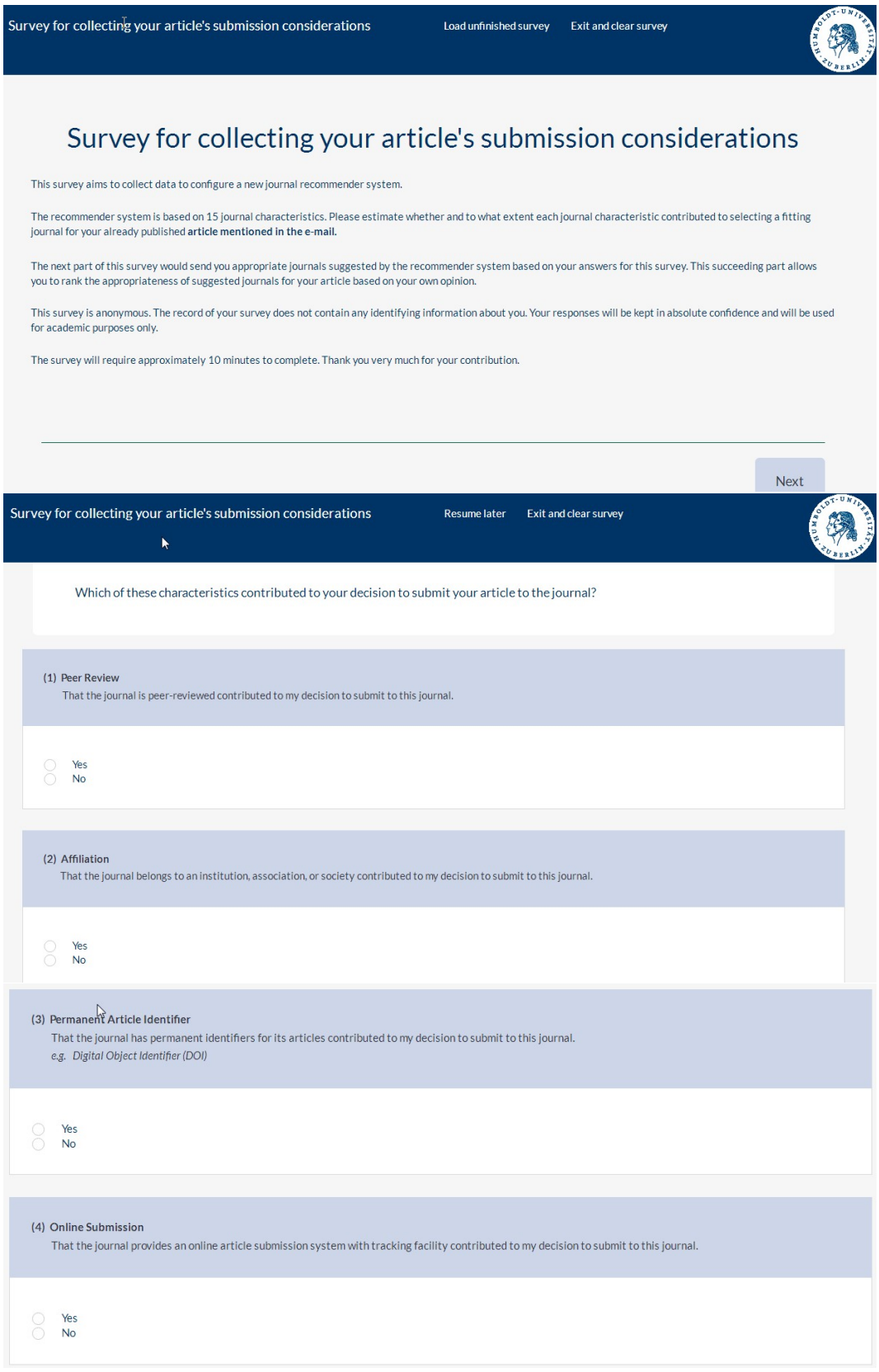
Thank you for taking time to assist me in my educational endeavors.

Sincerely,

P. K. C. M. Wijewickrema

Email: wijewicm@student.hu-berlin.de

C.2 Questionnaire



(5) Author Charges
That the journal is free of author charges (submission fee and article processing charge) contributed to my decision to submit to this journal.

- ○ Yes
- ● No

Please select the maximum amount of author charge (in US Dollar), which was acceptable to you when submitting the article.

Please choose...

(6) Abstracting and Indexing
That the journal is included in abstracting and indexing databases contributed to my decision to submit to this journal.

- ● Yes
- ○ No

Please select all abstracting and indexing databases, whose inclusion of the journal made you consider submitting to the journal.

- ☐ Academic Search Elite
- ☐ Applied Social Science Index and Abstracts (ASSIA)
- ☐ CAB Abstracts
- ☐ CINAHL
- ☐ Cochrane Library
- ☐ ERIC

(7) Journal's Prestige / Reputation
The journal's prestige / reputation level contributed to my decision to submit to this journal.

- ● Yes
- ○ No

Please select the (all) prestige / reputation level(s) of the journal, which was (were) acceptable to you when submitting the article.

e.g. if you considered journals with average or above average prestige / reputation, then please tick all 'Average', 'High', and 'Very high' options.

- ☐ Very low
- ☐ Low
- ☐ Average
- ☐ High
- ☐ Very high

(8) Publisher's Prestige / Reputation
The journal's publisher's prestige / reputation level contributed to my decision to submit to this journal.

- (●) Yes
- ( ) No

Please select the (all) prestige / reputation level(s) of the publisher, which was (were) acceptable to you when submitting the article.

- [ ] Very low
- [ ] Low
- [ ] Average
- [ ] High
- [ ] Very high

(9) Processing Time
The time it takes from submission to publication contributed to my decision to submit to this journal.

- (●) Yes
- ( ) No

Please select the (all) time span(s) from submission to publication, which was (were) acceptable to you when submitting the article.

- [ ] Very short
- [ ] Short
- [ ] Average
- [ ] Long
- [ ] Very long

(10) Acceptance Rate
The journal's acceptance rate contributed to my decision to submit to this journal.

- (●) Yes
- ( ) No

Please select the (all) level(s) of journal's acceptance rate(s), which was (were) acceptable to you when submitting the article.

- [ ] Very low
- [ ] Low
- [ ] Average
- [ ] High
- [ ] Very high

(11) Age
The journal's age contributed to my decision to submit to this journal.

- (●) Yes
- ( ) No

Please select the (all) age level(s) of the journal, which was (were) acceptable to you when submitting the article.

- [ ] Very recent
- [ ] Recent
- [ ] Middle-aged
- [ ] Old
- [ ] Very old

(12) Impact Factor
The journal's impact factor contributed to my decision to submit to this journal.

- (●) Yes
- ( ) No

Please select the (all) impact factor level(s) of the journal (relative to your subject domain), which was (were) acceptable to you when submitting the article.

- [ ] Very low
- [ ] Low
- [ ] Average
- [ ] High
- [ ] Very high

(13) Number of Issues
The number of issues of the journal per year contributed to my decision to submit to this journal.

- (●) Yes
- ( ) No

Please select the (all) range(s) for number of issues of the journal per year, which was (were) acceptable to you when submitting the article.

- [ ] 1-2 issues/year
- [ ] 3-5 issues/year
- [ ] 6-8 issues/year
- [ ] 9-12 issues/year
- [ ] Over 12 issues/year

(14) Number of Articles
The number of articles of the journal per issue contributed to my decision to submit to this journal.

- (●) Yes
- ( ) No

Please select the (all) range(s) for number of articles of the journal per issue, which was (were) acceptable to you when submitting the article.

- [ ] 1-10 articles/issue
- [ ] 11-30 articles/issue
- [ ] 31-60 articles/issue
- [ ] 61-100 articles/issue
- [ ] Over 100 articles/issue

(15) International Authorship
That the journal contains articles from authors from outside the country of origin of the journal contributed to my decision to submit to this journal.

- (●) Yes
- ( ) No

Please select the (all) range(s) for percentage of international authors per issue, which was (were) acceptable to you when submitting the article.

- [ ] 0-20%
- [ ] 21-40%
- [ ] 41-60%
- [ ] 61-80%
- [ ] 81-100%

Your further comments:

The final part of this survey will send you the suggested journals for your article based on the answers you provided. Would you be willing to participate in this final step?

- ( ) Yes
- ( ) No

The survey is now complete. Thank you once again for your time!

Please be assured that your responses will remain confidential. If you need additional information or have any questions, please feel free to contact me on wijewicm@student.hu-berlin.de

Appendix D

Third author survey: Evaluating the recommender system

D.1 Email invitation for third author survey

Subject: Results - Journal Recommendation

Dear Sir/Madam,

As a part of my PhD research at the Berlin School of Library and Information Science, Humboldt University of Berlin, Germany, I developed a novel journal recommender system, which could assist authors to select an appropriate journal for submitting manuscripts. This survey connects with the previous survey: "Survey for collecting your article's submission considerations", which was completed by you. The new journal recommender system has suggested 10 open access journals based on your answers for the previous survey and the text of your article. This final survey requests you to rank the appropriateness of suggested journals based on your own opinion.

The generated list is web-based. Please use the following link to view the list (length of survey: approximately 5 minutes):

https://docs.google.com/spreadsheets/d/1UwI78vZL1nLpuPcQFMiTyccO5lNGO3XkqW6c3l_sinA/edit?usp=sharing

Your responses will be kept in absolute confidence and will be used for academic purposes only. If you require additional information or have any questions, then please feel free to contact me.

Thank you for taking time to assist me in my educational endeavors.

Sincerely,

P. K. C. M. Wijewickrema

Email: wijewicm@student.hu-berlin.de

D.2 Questionnaire

Survey for collecting data to evaluate a journal recommender system

Table 1 shows the first 10 most appropriate open access journal titles suggested by the new recommender system for your article '**title of the article**'.

The rank order for suggested journals implies how far your article fit with each journal. The recommendations are based on two factors:

Factor 1: The text of your article.

Factor 2: Answers provided by you for the previous questionnaire. The previous questionnaire inquired whether and how you considered 15 journal characteristics while selecting an appropriate journal for your above mentioned article (see table 2 for your answers).

The first column of table 1 shows the journals suggested by the recommender system based on factors 1 and 2.

Please rank the journals according to how appropriate they are for your article in the **third column of table 1**. Use numbers from 1 to 10 to rank journals (1-most appropriate to 10-least appropriate). Indicate the rank as N/A, if a journal is not appropriate at all.

Your answers will be automatically saved to this form.

For your convenience, table 1 includes the original information of the characteristics of the suggested journals, while table 2 includes the answers you provided for the previous questionnaire (i.e. journal characteristics you expected).

Table 1: Suggested journals and their original information

Journal title	Order of appropriateness * (suggested by the recommender system)	Order of appropriateness (your opinion)	Subject(s) of journal	Peer reviewed	Belongs to an institution, association, or society	Has a permanent article identifier	Has an online submission system with tracking facility	Author charge	Journal's prestige	Publisher's prestige	Abstracted or indexed databases	Time from article submission to publication (~weeks)	Acceptance rate (~%)	Journal's age (~years)	~Impact factor	Number of issues (per year)	~Number of articles (per issue)	International authorship per issue (~%)
Title 01	1																	
Title 02	2																	
Title 03	3																	
Title 04	4																	
Title 05	5																	
Title 06	6																	
Title 07	7																	
Title 08	8																	
Title 09	9																	
Title 10	10																	

*** 1 to 10: most appropriate to least appropriate**

N/A: not appropriate at all

Table 2: Journal characteristics you expected for the article

Peer reviewed	Belongs to an institution, association, or society	Has a permanent article identifier	Has an online submission system with tracking facility	Author charge	Journal's prestige	Publisher's prestige	Abstracted or indexed databases	Time from article submission to publication (weeks)	Acceptance rate (%)	Journal's age (years)	Impact factor	Number of issues (per year)	Number of articles (per issue)	International authorship per issue (%)

Appendix E

First literature survey: Articles used for identifying factors

Following list provides the collection of previously published articles, which were used for identifying the important journal selection factors considered by the authors while selecting appropriate journal outlets.

1. Beaubien, S., and Eckard, M. (2014). Addressing faculty publishing concerns with open access journal quality indicators. *Journal of Librarianship and Scholarly Communication*, 2(2), p.eP1133. doi: 10.7710/2162-3309.1133. https://doi.org/10.7710/2162-3309.1133

2. Björk, B.-C., and Holmström, J. (2006). Benchmarking scientific journals from the submitting author's viewpoint. *Learned Publishing*, 19(2), 147—155. doi: 10.1087/095315106776387002. https://doi.org/10.1087/095315106776387002

3. Bröchner, J., and Björk, B. C. (2008). Where to submit? Journal choice by construction management authors. *Construction Management and Economics*, 26(7), 739—749. doi: 10.1080/01446190802017698. https://doi.org/10.1080/01446190802017698

4. Broome, M. E. (2007). A rose by any other name is still a rose: Assessing journal quality. *Nursing Outlook*, 55(4), 163-–164. doi: 10.1016/j.outlook.2007.06.001. https://doi.org/10.1016/j.outlook.2007.06.001

5. Butler, D. (2013). Investigating journals: The dark side of publishing. *Nature*, 495(7442), 433—435. doi: 10.1038/495433a.

6. Cotton, C. (2013). Submission fees and response times in academic publishing. *The American Economic Review*, 103(1), 501-–509. doi:10.2307/23469651 http://www.jstor.org/stable/23469651

7. Curry, M. J., and Lillis, T. M. (2010). Academic research networks: Accessing resources for English-medium publishing. *English for Specific Purposes*, 29(4), 281—295. doi: 10.1016/j.esp.2010.06.002. https://doi.org/10.1016/j.esp.2010.06.002

8. Curzan, A., and Queen, R. (2006). In the profession: Academic publication. *Journal of English Linguistics*, 34(4), 367-–372. doi: 10.1177/0075424206295808

9. Dalton, M. (2013). A dissemination divide? The factors that influence the journal selection decision of Library and Information Studies (LIS) researchers and practitioners. *Library and Information Research*, 37(115), 33-–57. doi:10.29173/lirg553. https://doi.org/10.29173/lirg553

10. Dalton, M. (2012). Traditional factors of fit, perceived quality, and speed of publication still outweigh open access in authors' journal selection criteria. *Evidence Based Library and Information Practice*, 7(4), 102—104. doi:10.18438/B8HC87. https://doi.org/10.18438/B8HC87

11. Gasparyan, A. Y. (2013). Choosing the target journal: Do authors need a comprehensive approach? *Journal of Korean Medical Science*, 28(8), 1117—1119. doi:10.3346/jkms.2013.28.8.1117. https://dx.doi.org/10.3346/jkms.2013.28.8.1117

12. Gutknecht, C. (2014). *Where to publish? Development of a recommender system for academic publishing* (Unpublished master thesis). University of Applied Sciences and Arts Northwestern Switzerland, Basel, Switzerland. http://hdl.handle.net/10760/23523

13. Heintzelman, M., and Nocetti, D. (2009). Where should we submit our manuscript? An analysis of journal submission strategies. *The B.E. Journal of Economic Analysis & Policy*, 9(1), 1-–28. https://EconPapers.repec.org/RePEc:bpj:bejeap:v:9:y:2009:i:1:n:39

14. Johnstone, M. J. (2007). Journal impact factors: Implications for the nursing profession. *International Nursing Review*, 54(1), 35-–40. doi:10.1111/j.1466-7657.2007.00527.x https://doi.org/10.1111/j.1466-7657.2007.00527.x

15. Klingner, J. K., Scanlon, D., and Pressley, M. (2005). How to publish in scholarly journals. *Educational Researcher*, 34(8), 14—20. doi:10.3102/0013189X034008014 https://doi.org/10.3102/0013189X034008014

16. Knight, L. V., Steinbach, T. A., and Levy, Y. (2008). Selecting an appropriate publication outlet: A comprehensive model of journal selection criteria for researchers in a broad range of academic disciplines. *International Journal of Doctoral Studies*, 3, 59—79.

17. Lewallen, L. P., and Crane, P. B. (2010). Choosing a publication venue. *Journal of Professional Nursing*, 26(4), 250—254. doi:10.1016/j.profnurs.2009.12.005 https://doi.org/10.1016/j.profnurs.2009.12.005

18. Milman, V. (2006). Impact factor and how it relates to quality of journals. *Notices of the American Mathematical Society*, 53(3), 351—352.

19. Özçakar, L., Franchignoni, F., Kara, M., and Muñoz, L. S. (2012). Choosing a scholarly journal during manuscript submission: The way how it rings true for physiatrists. *European Journal of Physical and Rehabilitation Medicine*, 48(4), 643-–647.

20. Peleg, R., and Shvartzman, P. (2006). Where should family medicine papers be published—Following the impact factor? *The Journal of the American Board of Family Medicine*, 19(6), 633-–636. doi:10.3122/jabfm.19.6.633. https://doi.org/10.3122/jabfm.19.6.633

21. Regazzi, J. J., and Aytac, S. (2008). Author perceptions of journal quality. *Learned Publishing*, 21(3), 225—235. doi:10.1087/095315108X288938. https://doi.org/10.1087/095315108X288938

22. Rowlands, I., and Nicholas, D. (2006). The changing scholarly communication landscape: An international survey of senior researchers. *Learned Publishing*, 19(1), 31—55. doi:10.1087/095315106775122493 https://doi.org/10.1087/095315106775122493

23. Rowlands, I., Nicholas, D., and Huntington, P. (2004). Scholarly communication in the digital environment: What do authors want? *Learned Publishing*, 17(4), 261—273. doi:10.1087/0953151042321680

24. Sharman, A. (2015). Where to publish. *The Annals of The Royal College of Surgeons of England*, 97(5), 329—332. doi:10.1308/rcsann.2015.0003 https://dx.doi.org/10.1308/rcsann.2015.0003

25. Shokraneh, F., Ilghami, R., Masoomi, R., and Amanollahi, A. (2012). How to select a journal to submit and publish your biomedical paper? *BioImpacts*, 2(1), 61–68. doi: 10.5681/bi.2012.008 https://dx.doi.org/10.5681/bi.2012.008

26. Solomon, D. J., and Björk, B. C. (2012). Publication fees in open access publishing: Sources of funding and factors influencing choice of journal. *Journal of the American Society for Information Science and Technology*, 63(1), 98-–107. doi: 10.1002/asi.21660 https://doi.org/10.1002/asi.21660

27. Stoilescu, D., and McDougall, D. (2010). Starting to publish academic research as a doctoral student. *International Journal of Doctoral Studies*, 5, 79—92. http://hdl.handle.net/1807/30064

28. Swan, A. (1999). 'What authors want': The ALPSP research study on the motivations and concerns of contributors to learned journals. *Learned Publishing*, 12(3), 170-–172. doi: 10.1087/09531519950145742 https://doi.org/10.1087/09531519950145742

29. Todd, R. W. (2014). Choosing venues for publishing research: A Thai perspective. *In Proceedings of the International Conference: DRAL 2/ILA 2014*, 37-–52. Thonburi, Bangkok: King Mongkut's University of Technology.

30. Uysal, H. H. (2012). The critical role of journal selection in scholarly publishing: A search for journal options in language-related research areas and disciplines. *Journal of Language and Linguistic Studies*, 8(1), 50—95.

31. van Teijlingen, E., and Hundley, V. (2002). Getting your paper to the right journal: A case study of an academic paper. *Journal of Advanced Nursing*, 37(6), 506—511. doi: 10.1046/j.1365-2648.2002.02135.x https://doi.org/10.1046/j.1365-2648.2002.02135.x

32. Warlick, S. E., and Vaughan, K. (2007). Factors influencing publication choice: Why faculty choose open access. *Biomedical Digital Libraries*, 4(1),1. doi: 10.1186/1742-5581-4-1 https://doi.org/10.1186/1742-5581-4-1

33. Welch, S. J. (2012). Selecting the right journal for your submission. *Journal of Thoracic Disease*, 4(3), 336-–338. doi: 10.3978/j.issn.2072-1439.2012.05.06 http://jtd.amegroups.com/article/view/389

34. Wijewickrema, P. K. C. M. (2015). A three dimensional model for selecting the most appropriate journal outlet for manuscript submission. *Paper presented at the International Conference on Social Science Research, ICSSR 2015*, Malaysia.

35. Witt, P. A. (2003). Readership is more important than publication outlet. *Journal of Leisure Research*, 35(3), 331—334. doi: 10.1080/00222216.2003.11949999 https://doi.org/10.1080/00222216.2003.11949999

36. Ziobrowski, A., and Gibler, K. (2000). Factors academic real estate authors consider when choosing where to submit a manuscript for publication. *Journal of Real Estate Practice and Education*, 3(1), 43—54. doi: 10.5555/repe.3.1.1762151051km2227 https://www.aresjournals.org/doi/pdf/10.5555/repe.3.1.1762151051km2227

Appendix F

Second literature survey: Articles used for identifying A&I services

Following list provides the collection of previously published articles, which were studied for identifying the most influential abstracting and indexing services for medicine and social sciences journals.

1. Anderson, J. D., and Pérez-Carballo, J. (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort. *Information Processing & Management*, 37(2), 255-277. doi: 10.1016/S0306-4573(00)00046-7.

2. Archambault, É., Campbell, D., Gingras, Y., and Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the Association for Information Science and Technology*, 60(7), 1320-1326. doi: 10.1002/asi.21062. https://doi.org/10.1002/asi.21062

3. Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M., and Rogers, W. J. (2004). The NLM indexing initiative's medical text indexer. *Medinfo. Part 2 of Medinfo 2004: Proceedings of the 11th World Congress on Medical Informatics*, 89, 268–272.

4. Asche, C., LaFleur, J., and Conner, C. (2011). A review of diabetes treatment adherence and the association with clinical and economic outcomes. *Clinical thera-*

peutics, 33(1), 74-109. doi: 10.1016/j.clinthera.2011.01.019. https://doi.org/10.1016/j.clinthera.2011.01.019

5. Baker, P. R., Francis, D. P., Soares, J., Weightman, A. L., and Foster, C. (2011). Community wide interventions for increasing physical activity. *Sao Paulo Medical Journal*, 129(6), 436–437. doi: 10.1590/S1516-31802011000600013. https://dx.doi.org/10.1590/S1516-31802011000600013

6. Bakkalbasi, N., Bauer, K., Glover, J., and Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical digital libraries*, 3(1), 7. doi :10.1186/1742-5581-3-7. https://doi.org/10.1186/1742-5581-3-7

7. Ball, R., and Tunger, D. (2006). Science indicators revisited–Science Citation Index versus SCOPUS: A bibliometric comparison of both citation databases. *Information Services and Use*, 26(4), 293–301. doi: 10.3233/ISU-2006-26404

8. Bar-Ilan, J. (2008). Which h-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), 257–271. doi: 10.1007/s11192-008-0216-y. https://doi.org/10.1007/s11192-008-0216-y

9. Bar-Oz, B., Einarson, T., Einarson, A., Boskovic, R., O'Brien, L., Malm, H., Bérard, A. and Koren, G. (2007). Paroxetine and congenital malformations: meta-analysis and consideration of potential confounding factors. *Clinical therapeutics*, 29(5), 918–926. doi: 10.1016/j.clinthera.2007.05.003. https://doi.org/10.1016/j.clinthera.2007.05.003

10. Björk, B. C., and Solomon, D. (2012). Open access versus subscription journals: a comparison of scientific impact. *BMC medicine*, 10(1), 73. doi: 10.1186/1741-7015-10-73. https://doi.org/10.1186/1741-7015-10-73

11. Burnham, J. F. (2006). Scopus database: a review. *Biomedical digital libraries*, 3(1), 1. doi: 10.1186/1742-5581-3-1. https://doi.org/10.1186/1742-5581-3-1

12. Callaham, M., Wears, R. L., and Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA*, 287(21), 2847–2850. doi:10.1001/jama.287.21.2847.

13. Chadegani, A. A., Salehi, H., Yunus, M., Farhadi, H., Fooladi, M., Farhadi, M., and Ale Ebrahim, N. (2013). A comparison between two main academic literature collections: Web of Science and Scopus databases. *Asian Social Science*, 9(5), 18–26. doi: 10.5539/ass.v9n5p18. http://dx.doi.org/10.5539/ass.v9n5p18

14. Corley, D. A., Kerlikowske, K., Verma, R., and Buffler, P. (2003). Protective association of aspirin/NSAIDs and esophageal cancer: a systematic review and meta-analysis. *Gastroenterology*, 124(1), 47–56. https://doi.org/10.1053/gast.2003.50008

15. Coulter, A., and Ellins, J. (2007). Effectiveness of strategies for informing, educating, and involving patients. *BMJ: British Medical Journal*, 335(7609), 24–27. doi: 10.1136/bmj.39246.581169.80. https://doi.org/10.1136/bmj.39246.581169.80

16. Cronin, B. (2001). Bibliometrics and beyond: some thoughts on web-based citation analysis. *Journal of Information science*, 27(1), 1–7. https://doi.org/10.1177/016555150102700101

17. de Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F. J., González-Molina, A., and Herrero-Solana, V. (2007). Coverage analysis of Scopus: A journal metric approach. *Scientometrics*, 73(1), 53–78. doi: 10.1007/s11192-007-1681-4. https://doi.org/10.1007/s11192-007-1681-4

18. Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., and Pappas, G. (2008). Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. *The FASEB journal*, 22(2), 338–342. doi: 10.1096/fj.07-9492LSF. https://doi.org/10.1096/fj.07-9492LSF

19. Fangerau, H. (2004). Finding European bioethical literature: an evaluation of the leading abstracting and indexing services. *Journal of medical ethics*, 30(3), 299–303. doi: 10.1136/jme.2003.003269. http://dx.doi.org/10.1136/jme.2003.003269

20. Franceschini, F., Maisano, D., and Mastrogiacomo, L. (2016). The museum of errors/horrors in Scopus. *Journal of Informetrics*, 10(1), 174–182. doi: 10.1016/j.joi.2015.11.006. https://doi.org/10.1016/j.joi.2015.11.006

21. Garrard, J. (2013). Health sciences literature review made easy. Jones and Bartlett Publishers.

22. Giangregorio, L., Papaioannou, A., Cranney, A., Zytaruk, N., and Adachi, J. D. (2006, April). Fragility fractures and the osteoporosis care gap: an international phenomenon. *Seminars in arthritis and rheumatism*, 35(5), 293–305. doi: 10.1016/j.semarthrit.2005.11.001. https://doi.org/10.1016/j.semarthrit.2005.11.001

23. Giustini, D., and Barsky, E. (2005). A look at Google Scholar, PubMed, and Scirus: comparisons and recommendations. *Journal of the Canadian Health Libraries Association*, 26(3), 85–89. doi: 10.5596/c05-030. https://doi.org/10.5596/c05-030

24. Glanville, J. M., Lefebvre, C., Miles, J. N., and Camosso-Stefinovic, J. (2006). How to identify randomized controlled trials in MEDLINE: ten years on. *Journal of the Medical Library Association*, 94(2), 130–136.

25. Greenhalgh, T., and Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ*, 331(7524), 1064–1065. doi: 10.1136/bmj.38636.593461.68. https://doi.org/10.1136/bmj.38636.593461.68

26. Grindlay, D. J., Brennan, M. L., and Dean, R. S. (2012). Searching the veterinary literature: a comparison of the coverage of veterinary journals by nine bibliographic databases. *Journal of veterinary medical education*, 39(4), 404–412. doi: 10.3138/jvme.1111.109R. https://doi.org/10.3138/jvme.1111.109R

27. Groll, D. L., To, T., Bombardier, C., and Wright, J. G. (2005). The development of a comorbidity index with physical function as the outcome. *Journal of clinical epidemiology*, 58(6), 595–602. doi: 10.1016/j.jclinepi.2004.10.018. https://doi.org/10.1016/j.jclinepi.2004.10.018

28. Hjørland, B. (2002). Domain analysis in information science: eleven approaches–traditional as well as innovative. *Journal of documentation*, 58(4), 422–462. doi: 10.1108/00220410210431136. https://doi.org/10.1108/00220410210431136

29. Hsu, P. P. (2002). Natural medicines comprehensive database. *Journal of the Medical Library Association*, 90(1), 113–114.

30. Jacsó, P. (2004). Citation-enhanced indexing/abstracting databases. *Online information review*, 28(3), 235–238. doi: 10.1108/14684520410543689. https://doi.org/10.1108/14684520410543689

31. Jacso, P. (2005a). As we may search—comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current science*, 89(9), 1537–1547. http://www.jstor.org/stable/24110924

32. Jacsó, P. (2005b). Comparison and analysis of the citedness scores in Web of Science and Google Scholar. *Digital Libraries: Implementing Strategies and Sharing Experiences*, 3815, 360–369.

33. Jacsó, P. (2008). Google scholar revisited. *Online information review*, 32(1), 102–114. doi: 10.1108/14684520810866010. https://doi.org/10.1108/14684520810866010

34. Jacsó, P. (2011). The h-index, h-core citation rate and the bibliometric profile of the Scopus database. *Online Information Review*, 35(3), 492–501. doi: 10.1108 /14684521111151487. https://doi.org/10.1108/14684521111151487

35. Klavans, R., and Boyack, K. W. (2007). Is there a convergent structure of science? A comparison of maps using the ISI and Scopus databases. *Proceedings of ISSI*, 1, 437–448.

36. Kousha, K., and Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, 74(2), 273–294. doi=10.1007/s11192-008-0217-x. https://doi.org/10.1007/s11192-008-0217-x

37. Kulkarni, A. V., Aziz, B., Shams, I., and Busse, J. W. (2009). Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *JAMA*, 302(10), 1092–1096. doi: 10.1001/jama.2009.1307.

38. Larsen, P. O., and von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3), 575–603. doi:10.1007/s11192-010-0202-z. https://doi.org/10.1007/s11192-010-0202-z

39. Legg, Lynn and Drummond, Avril and Leonardi-Bee, Jo and Gladman, J R F and Corr, Susan and Donkervoort, Mireille and Edmans, Judi and Gilbertson, Louise and Jongbloed, Lyn and Logan, Pip and Sackley, Catherine and Walker, Marion and Langhorne, Peter Legg, L., Drummond, A., Leonardi-Bee, J., Gladman, J. R. F., Corr, S., Donkervoort, M., Edmans, J., Gilbertson, L., Jongbloed, L., Logan, P., Sackley, C., Walker, M., and Langhorne, P. (2007). Occupational therapy for patients with problems in personal activities of daily living after stroke: systematic review of randomised trials. *BMJ*, 335(7626), 922. doi:10.1136/bmj.39343.466863.55. https://doi.org/10.1136/bmj.39343.466863.55

40. Mair, F., and Whitten, P. (2000). Systematic review of studies of patient satisfaction with telemedicine. *BMJ*, 320(7248), 1517–1520. doi: 10.1136/bmj.320.7248.1517. https://doi.org/10.1136/bmj.320.7248.1517

41. McDonald, S., Taylor, L., and Adams, C. (1999). Searching the right database. A comparison of four databases for psychiatry journals. *Health Information and Libraries Journal*, 16(3), 151–156. doi:j.1365-2532.1999.00222.x. https://doi.org/10.1046/j.1365-2532.1999.00222.x

42. Meho, L. I., and Tibbo, H. R. (2003). Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the Association for Information Science and Technology*, 54(6), 570–587. doi: 10.1002/asi.10244. https://doi.org/10.1002/asi.10244

43. Meho, L. I., and Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58(13), 2105–2125. doi: 10.1002/asi.20677. https://doi.org/10.1002/asi.20677

44. Micha, R., Imamura, F., von Ballmoos, M. W., Solomon, D. H., Hernán, M. A., Ridker, P. M., and Mozaffarian, D. (2011). Systematic review and meta-analysis of methotrexate use and risk of cardiovascular disease. *American Journal of Cardiology*, 108(9), 1362–1370. doi: 10.1016/j.amjcard.2011.06.054. https://doi.org/10.1016/j.amjcard.2011.06.054

45. Mor Barak, M. E., Travis, D. J., Pyun, H., and Xie, B. (2009). The impact of supervision on worker outcomes: A meta-analysis. *Social Service Review*, 83(1), 3–32.

46. Norris, M., and Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of informetrics*, 1(2), 161–169. doi: 10.1016/j.joi.2006.12.001. https://doi.org/10.1016/j.joi.2006.12.001

47. Noruzi, A. (2005). Google Scholar: The new generation of citation indexes. *LIBRI*, 55(4), 170–180.

48. Notess, G.R. (2005). Scholarly web searching: Google Scholar and Scirus. *Online*, 29(4), 39–41.

49. Orcher, L. T. (2016). Conducting research: Social and behavioral science methods. Routledge.

50. Otte, E., and Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science*, 28(6), 441–453. doi: 10.1177/016555150202800601. https://doi.org/10.1177/016555150202800601

51. Pan, C. X., Morrison, R. S., Ness, J., Fugh-Berman, A., and Leipzig, R. M. (2000). Complementary and alternative medicine in the management of pain, dyspnea, and nausea and vomiting near the end of life: a systematic review. *Journal of pain*

and symptom management, 20(5), 374–387. doi: 10.1016/S0885-3924(00)00190-1. https://doi.org/10.1016/S0885-3924(00)00190-1

52. Prosser, I., Maguire, S., Harrison, S. K., Mann, M., Sibert, J. R., and Kemp, A. M. (2005). How old is this fracture? Radiologic dating of fractures in children: a systematic review. *American Journal of Roentgenology*, 184(4), 1282–1286. doi: 10.2214/ajr.184.4.01841282. https://doi.org/10.2214/ajr.184.4.01841282

53. Royle, P., and Milne, R. (2003). Literature searching for randomized controlled trials used in Cochrane reviews: rapid versus exhaustive searches. *International journal of technology assessment in health care*, 19(4), 591–603. doi: 10.1017/S0266462303000552. https://doi.org/10.1017/S0266462303000552

54. Sadeh, A. (2011). The role and validity of actigraphy in sleep medicine: an update. *Sleep medicine reviews*, 15(4), 259–267. doi: 10.1016/j.smrv.2010.10.001. https://doi.org/10.1016/j.smrv.2010.10.001

55. Scherer, R. W., Langenberg, P., and Von Elm, E. (2007). Full publication of results initially presented in abstracts. *Cochrane Database of Systematic Reviews*, 2007(2). doi: 10.1002/14651858.MR000005.pub3. https://doi.org/10.1002/14651858.MR000005.pub3

56. Schultz, M. (2007). Comparing test searches in PubMed and Google Scholar. *Journal of the Medical Library Association*, 95(4), 442–445. doi: 10.3163/1536-5050.95.4.442. https://dx.doi.org/10.3163/1536-5050.95.4.442

57. Shaw, R. L., Booth, A., Sutton, A. J., Miller, T., Smith, J. A., Young, B., Jones, D.R., and Dixon-Woods, M. (2004). Finding qualitative research: an evaluation of search strategies. *BMC medical research methodology*, 4(1),5. doi: 10.1186/1471-2288-4-5. https://doi.org/10.1186/1471-2288-4-5

58. Singer, G. H. (2006). Meta-analysis of comparative studies of depression in mothers of children with and without developmental disabilities. *American journal on mental retardation*, 111(3), 155–169. https://doi.org/10.1352/0895-8017(2006)111[155:MOCSOD]2.0.CO;2

59. Suleymenov, E. Z., Ponomareva, N. I., Dzhumabekov, A. K., Kubieva, T. S., and Kozbagarova, G. A. (2011). An assessment of the contributions of Kazakhstan and other CIS countries to global science: the Scopus database. *Scientific and Technical*

Information Processing, 38(3), 159–165. doi: 10.3103/S0147688211030051. https://doi.org/10.3103/S0147688211030051

60. Sweileh, W. M., Shraim, N. Y., Al-Jabi, S. W., Sawalha, A. F., Rahhal, B., Khayyat, R. A., and Sa'ed, H. Z. (2016). Assessing worldwide research activity on probiotics in pediatrics using Scopus database: 1994—2014. *World Allergy Organization Journal*, 9(1), 25. doi: 10.1186/s40413-016-0116-1. https://doi.org/10.1186/s40413-016-0116-1

61. Talja, S., and Maula, H. (2003). Reasons for the use and non-use of electronic journals and databases: A domain analytic study in four scholarly disciplines. *Journal of Documentation*, 59(6), 673–691. doi: 10.1108/00220410310506312. https://doi.org/10.1108/00220410310506312

62. Trifiletti, L. B., Gielen, A. C., Sleet, D. A., and Hopkins, K. (2005). Behavioral and social sciences theories and models: are they used in unintentional injury prevention research?. *Health Education Research*, 20(3), 298–307. doi: 10.1093/her/cyg126. https://doi.org/10.1093/her/cyg126

63. Vibert, N., Ros, C., Bigot, L. L., Ramond, M., Gatefin, J., and Rouet, J. F. (2009). Effects of domain knowledge on reference search with the PubMed database: An experimental study. *Journal of the Association for Information Science and Technology*, 60(7), 1423–1447. doi: 10.1002/asi.21078. https://doi.org/10.1002/asi.21078

64. Walters, W.H. (2007). Google Scholar coverage of a multidisciplinary field. *Information Processing and Management*, 43(4), 1121–1132. doi: 10.1016/j.ipm.2006.08.006.

65. Yang, K., and Meho, L.I. (2006). Citation analysis: a comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the 69th Annual Meeting of the American Society for Information Science and Technology (ASIST)*, 43(1), 1–15.

Appendix G

Corpora of journals

Following journal lists represent all the journal titles used for constructing the two training journal corpora – medicine and social sciences of the current study.

No.	Medicine	Social sciences
001	Academia Anatomica International	3L The Southeast Asian Journal of English Language Studies
002	Acta Angiologica	A Contracorriente
003	Acta Medica	Academic journal of business, administration, law and social sciences
004	Acta Medica Bulgarica	Academicus International scientific journal
005	Acta Medica Marisiensis	ACE Architecture, City and Environment
006	Acta Medica Martiniana	Acta Geographica Debrecina
007	Acta Médica Portuguesa	Acta Universitaria
008	Acta Ortopédica Brasileira	Acta Universitatis Danubius. Oeconomica
009	Advanced Biomedical Research	Acta Universitatis Sapientiae, Social Analyses

continued . . .

. . . continued

No.	Medicine	Social sciences
010	Advances in Bioscience and Clinical Medicine	Adeptus
011	Advances in Epidemiology	AD-minister
012	Advances in Interventional Cardiology	Advances in Language and Literary Studies
013	Advances in Respiratory Medicine	AERA Open
014	African Journal of Disability	Africa Spectrum
015	African Journal of Emergency Medicine	African Journal of Disability
016	African Journal of Health Professions Education	African Journal of Health Professions Education
017	African Journal of Paediatric Surgery	Agronomy
018	African Journal of Primary Health Care and Family Medicine	Akademika
019	Althea Medical Journal	Akroterion
020	American Journal of Experimental and Clinical Research	Al-Iqtishad Jurnal Ilmu Ekonomi Syariah
021	Ancient Science of Life	Al-Ta'lim Journal
022	Angiologia e Cirurgia Vascular	Alternate routes Journal of critical social research
023	Ankara Medical Journal	Ambiances
024	Annals of Cardiac Anaesthesia	American studies journal
025	Annals of Eurasian Medicine	Analele Universitatii Constantin Brancusi din Targu Jiu Seria Economie
026	Annals of Indian Academy of Neurology	Analisa Journal of Social Science and Religion
027	Annals of Pediatric Cardiology	Anglophonia
028	Annals of Tropical Medicine and Public Health	Annals of Agricultural and Environmental Medicine
029	Archives of Biomedical Sciences	Annals of Applied Sport Science

continued . . .

... continued

No.	Medicine	Social sciences
030	Archives of Endocrinology and Metabolism	Annals of Philosophy, Social and Human Disciplines
031	ARS Medica Tomitana	Annals of the University of Bucharest Geography Series
032	Arthritis Research and Therapy	Antrocom Journal of Anthropology
033	Asia Pacific Allergy	Apparatus
034	Asian Pacific Journal of Reproduction	Applied Research in Health and Social Sciences
035	Asian Pacific Journal of Tropical Disease	Apuntes Revista de Ciencias Sociales
036	Autopsy and Case Reports	Arabian Epigraphic Notes
037	Bali Medical Journal	Arctic and North
038	Balkan Medical Journal	Asean Marketing Journal
039	Bangabandhu Sheikh Mujib Medical University Journal	ASEAS
040	Bangladesh Journal of Medical Science	Asia and the Pacific Policy Studies
041	Bengal Journal of Otolaryngology and Head Neck Surgery	Asia Pacific Journal of Multidisciplinary Research
042	Bioautomation	Atlantic Review of Economics
043	Biology of Sex Differences	Australian Educational Computing
044	Biomedical Glasses	Australian Journal of Business Management Research
045	Biomedical Human Kinetics	Australian Review of Public Affairs
046	BioResearch Open Access	Austrian Journal of Political Science
047	Birat Journal of Health Sciences	Authorship
048	Blood Transfusion	Barents Studies Peoples, Economies and Politics
049	BMC Hematology	Bearing Witness Joyce Carol Oates Studies
050	BMC Medicine	Behemoth a Journal on Civilisation

continued ...

. . . continued

No.	Medicine	Social sciences
051	BMC Proceedings	Bellaterra Journal of Teaching and Learning Language and Literature
052	BMC Research Notes	Berkeley Planning Journal
053	BMJ Open	Bilgi Dunyasi
054	BMJ Open Respiratory Research	Binus Business Review
055	Brazilian Journal of Forensic Sciences, Medical Law and Bioethics	BioethiqueOnline
056	Brazilian Journal of Medicine and Human Health	Biolinguistics
057	Brazilian Neurosurgery	Biology of Exercise
058	Burns and Trauma	Boletim do Museu Paraense Emílio Goeldi
059	Cadernos de Saúde Pública	Boreal Environment Research
060	Cardiology and Therapy	Boston Hospitality Review
061	Cardiology Journal	Brazilian Journal of Science and Technology
062	Cardiometry	Brno Studies in English
063	Cell Communication and Signaling	Brussels Studies
064	Central European Journal of Immunology	Bulgarian Journal of Science and Education Policy
065	Central European Journal of Urology	Canadian Journal for the Scholarship of Teaching and Learning
066	Chinese Journal of Cancer	Canadian Journal of Higher Education
067	Chinese Medical Journal	Canadian Journal of Learning and Technology
068	Chiropractic and Osteopathy	Canadian Journal of Nonprofit and Social Economy Research
069	Clinical and Laboratorial Research in Dentistry	Canadian Studies in Population
070	Clinical Case Reports	Catholic Social Science Review

continued . . .

... continued

No.	Medicine	Social sciences
071	Clinical Epidemiology	Central European Journal of Public Policy
072	Clinical Epigenetics	Chinese Librarianship an International Electronic Journal
073	Clinical Hypertension	Cinema Journal of Philosophy and the Moving Image
074	Clinical Medicine Insights Trauma and Intensive Medicine	City Territory and Architecture
075	Clinical Phytoscience	CLCWeb
076	Clinical Proteomics	CLEaR
077	Contemporary Clinical Dentistry	Človek a Spoločnosť
078	Contemporary Oncology	Code4Lib Journal
079	Crescent Journal of Medical and Biological Sciences	Cogent Arts and Humanities
080	Critical Care	Cogent Business and Management
081	Current Directions in Biomedical Engineering	Cogent Social Sciences
082	Current Issues in Pharmacy and Medical Sciences	Collabra
083	Current Pediatric Research	College and Research Libraries
084	Dalhousie Medical Journal	Communications in Information Literacy
085	Dataset Papers in Science	Comparative Migration Studies
086	Delta Medical College Journal	Comunicar
087	Dermatology and Therapy	Conservar Património
088	Dermatology Review	Conservation and Society
089	Diabetes Therapy	Consilience The Journal of Sustainable Development
090	Dialogues in Philosophy, Mental and Neuro Sciences	Contemporary Economy

continued ...

... continued

No.	Medicine	Social sciences
091	Disease and Molecular Medicine	Contemporary Issues in Technology and Teacher Education
092	Disease Models and Mechanisms	Contemporary Southeastern Europe
093	Diseases	Corvinus Journal of Sociology and Social Policy
094	EAI Endorsed Transactions on Pervasive Health and Technology	Creativity and Innovation Journal
095	EBioMedicine	Critical Questions in Education
096	Ecancermedicalscience	Croatian Economic Survey
097	Einstein (São Paulo)	Cromohs
098	EJNMMI Physics	Cuadernos de Economía
099	Electronic Journal of Health Informatics	Czech Journal of Tourism
100	eLife	Data
101	Emerging Infectious Diseases	Decyzje
102	Endokrynologia Polska	Delaware Review of Latin American Studies
103	Epidemiologia e Serviços de Saúde	Design and Technology Education an International Journal
104	ePlasty	Designs for Learning
105	ERJ Open Research	Developing Country Studies
106	European Clinical Respiratory Journal	Digital Culture and Education
107	European Journal of Bioethics	Digital Defoe
108	European Journal of Case Reports in Internal Medicine	Digital Education Review
109	European Journal of General Dentistry	Digithum
110	European Journal of Medical Research	Discobolul
111	European Journal of Translational Myology	Dubrovnik Annals
112	European Medical Journal Diabetes	Earth Perspectives

continued ...

... continued

No.	Medicine	Social sciences
113	European Medical Journal Gastroenterology	Earth, Planets and Space
114	European Medical Journal Neurology	Eastern Journal of European Studies
115	European Medical Journal Oncology	Economia Aziendale Online
116	European Medical Journal Rheumatology	Economic Insights – Trends and Challenges
117	European Medical Journal Urology	Economics Management Innovation
118	European Pharmaceutical Journal	Economics The open-access Open-assessment e-journal
119	European Review of Aging and Physical Activity	Economy and Sociology
120	F1000Research	Economy Transdisciplinarity Cognition
121	Family Medicine and Primary Care Review	Education Policy Analysis Archives
122	Farmacia Hospitalaria	Education Sciences
123	Fertility Research and Practice	Educational Process International Journal
124	Fisioterapia em Movimento	Educational Research in Medical Sciences Journal
125	Folia Cardiologica	eJournal of eDemocracy and Open Government
126	Folia Medica	E-Journal of Tourism
127	Folia Medica Copernicana	Electronic journal of business ethics and orgnization studies
128	Folia Medica Facultatis Medicinae Universitatis Saraeviensis	Electronic journal of contemporary Japanese studies
129	Folia Neuropathologica	Electronic Journal of e-Learning
130	Frontiers in Neuroenergetics	Electronic Journal of Foreign Language Teaching
131	Future Science OA	electronic Journal of Health Informatics

continued ...

. . . continued

No.	Medicine	Social sciences
132	Galicia Clínica	Electronic Journal of Knowledge Management
133	Galician Medical Journal	Enhancing the Learner Experience in Higher Education
134	Gastroenterology Research	Enlightening Tourism A Pathmaking Journal
135	Gastroenterology Review	Entrepreneurial Business and Economics Review
136	Gazi Medical Journal	EnvironmentAsia
137	Genome Medicine	Ephemera
138	GMS Hygiene and Infection Control	Estudios Irlandeses
139	GMS Interdisciplinary Plastic and Reconstructive Surgery DGPW	ETHOS
140	GMS Journal for Medical Education	Études Caribéennes
141	Health Economics Review	Eurasian Journal of Economics and Finance
142	Health Problems of Civilization	Eurasian Journal of Social Sciences
143	Health Professional Student Journal	EuroEconomica
144	Health Psychology and Behavioral Medicine	European Countryside
145	Health Psychology Report	European Integration Studies
146	Health Psychology Research	European Journal of American Studies
147	Health Risk Analysis	European Journal of Business and Economics
148	Healthcare	European Journal of Government and Economics
149	Heart Views	European Journal of Social and Behavioural Sciences
150	HIV and AIDS Review. International Journal of HIV-Related Problems	European Journal of Sustainable Development

continued . . .

... continued

No.	Medicine	Social sciences
151	HIVAIDS Research and Palliative Care	European Quarterly of Political Attitudes and Mentalities
152	HNE Handover For Nurses and Midwives	European Researcher Series A
153	Hong Kong Medical Journal	Europolity Continuity and Change in European Governance
154	Human Genomics	Evidence Based Library and Information Practice
155	Immunopathologia Persa	Facta Universitatis, Series
156	Indian Journal of Anaesthesia	Fashion and Textiles
157	Indian Journal of Cancer	Fast Capitalism
158	Indian Journal of Community Health	Field Actions Science Report
159	Indian Journal of Community Medicine	Film Criticism
160	Indian Journal of Critical Care Medicine	Financial Theory and Practice
161	Indian Journal of Dental Research	Folia Oeconomica Stetinensia
162	Indian Journal of Dermatology	Forum
163	Indian Journal of Dermatology, Venereology and Leprology	Forum geographic
164	Indian Journal of Medical and Paediatric Oncology	Future Studies Research Journal
165	Indian Journal of Medical Research	Games
166	Indian Journal of Neonatal Medicine and Research	Gandhara Journal of Research in Social Science
167	Indian Journal of Occupational and Environmental Medicine	GEMA Online Journal of Language Studies
168	Indian Journal of Ophthalmology	Gender Forum
169	Indian Journal of Orthopaedics	Gender Studies
170	Indian Journal of Palliative Care	Geoenvironmental Disasters

continued ...

... continued

No.	Medicine	Social sciences
171	Indian Journal of Pathology and Microbiology	GeoJournal of Tourism and Geosites
172	Indian Journal of Pharmacology	Geoplanning Journal of Geomatics and Planning
173	Indian Journal of Plastic Surgery	Global Advances in Business Communication
174	Indian Journal of Public Health	Global Economic Observer
175	Indian Journal of Radiology and Imaging	Gymnasium
176	Indian Journal of Sexually Transmitted Diseases	Health and Justice
177	Infectious Diseases and Therapy	Health Professional Student Journal
178	Inflammation and Cell Signaling	Hellenic Journal of Music, Education, and Culture
179	Influenza and Other Respiratory Viruses	Heritage Science
180	Injury Epidemiology	Higher Learning Research Communications
181	Inside the Cell	Historical Review
182	Insights into Imaging	HOW
183	Intensive Care Medicine Experimental	Hydrology and Earth System Sciences
184	International Archives of Otorhinolaryngology	ILIRIA International Review
185	International Journal of Anatomical Variations	Im@go
186	International Journal of Biomedicine	In the Library with the Lead Pipe
187	International Journal of Brain Science	Indonesian Capital Market Review
188	International Journal of Cancer Therapy and Oncology	Indonesian EFL Journal

continued ...

... continued

No.	Medicine	Social sciences
189	International Journal of Clinical Transfusion Medicine	Indonesian Journal of Educational Review
190	International Journal of Collaborative Research on Internal Medicine and Public Health	Industrija
191	International Journal of Community Based Nursing and Midwifery	Informatica Economica
192	International Journal of Environmental Research and Public Health	Information for Social Change
193	International Journal of Health and Allied Sciences	Information Management and Business Review
194	International Journal of Health and Rehabilitation Sciences	Information Research
195	International Journal of Hepatobiliary and Pancreatic Diseases	Information Technology and Libraries
196	International Journal of Implant Dentistry	Informing Science The International Journal of an Emerging Transdiscipline
197	International Journal of Integrated Health Sciences	Infrastructure Complexity
198	International Journal of Medical Reviews	InMedia
199	International Journal of Medical Sciences	Innovar
200	International Journal of Medicine and Medical Research	InSight A Journal of Scholarly Teaching
201	International Journal of Occupational Medicine and Environmental Health	Insights
202	International Journal of One Health	Integral Review
203	International Journal of Otolaryngology	Interdisciplinary Description of Complex Systems

continued ...

. . . continued

No.	Medicine	Social sciences
204	International Journal of Palliative Care	Interdisciplinary Journal of e-Skills and Lifelong Learning
205	International Journal of Phytopharmacy	Interdisciplinary Journal of Information, Knowledge, and Management
206	International Journal of Preventive Medicine	Interface a Journal for and about Social Movements
207	International Journal of Pure and Applied Sciences and Technology	International and Multidisciplinary Journal of Social Sciences
208	International Journal of Shoulder Surgery	International Business and Economics Research Journal
209	International Journal of Surgery and Medicine	International Development Policy
210	International Journal of Telemedicine and Applications	International Educational E-Journal
211	International Journal of Whole Person Care	International Electronic Journal of Elementary Education
212	International Journal of Yoga	International Indigenous Policy Journal
213	International Maritime Health	International Journal for Crime, Justice and Social Democracy
214	International Medical Journal Malaysia	International Journal for the Scholarship of Teaching and Learning
215	International Practice Development Journal	International Journal for Transformative Research
216	Internet Journal of Medical Update	International Journal of Advances in Management and Economics
217	Iranian Journal of Basic Medical Sciences	International Journal of Area Studies
218	Iranian Journal of Neurology	International Journal of Assessment Tools in Education

continued . . .

. . . continued

No.	Medicine	Social sciences
219	ISRN Geriatrics	International Journal of Bahamian Studies
220	ISRN Hematology	International Journal of Conflict and Violence
221	ISRN Hepatology	International Journal of Conservation Science
222	ISRN Immunology	International Journal of Contemporary Educational Research
223	ISRN Infectious Diseases	International Journal of Digital Curation
224	ISRN Inflammation	International Journal of Digital Library Services
225	ISRN Nephrology	International Journal of Disaster Risk Science
226	ISRN Neuroscience	International Journal of Doctoral Studies
227	ISRN Nursing	International Journal of Economics and Financial Issues
228	ISRN Obesity	International Journal of Education and the Arts
229	ISRN Oncology	International Journal of Education and Literacy Studies
230	ISRN Ophthalmology	International Journal of English Language and Translation Studies
231	ISRN Orthopedics	International Journal of Finance and Banking Studies
232	ISRN Pain	International Journal of Information Dissemination and Technology
233	ISRN Parasitology	International Journal of Information Science and Management

continued . . .

... continued

No.	Medicine	Social sciences
234	ISRN Pathology	International Journal of Innovation
235	ISRN Pharmaceutics	International Journal of Instruction
236	ISRN Pharmacology	International Journal of Islamic Economics and Finance Studies
237	ISRN Plastic Surgery	International Journal of Knowledge Content Development and Technology
238	ISRN Psychiatry	International Journal of Languages' Education and Teaching
239	ISRN Pulmonology	International Journal of Lean Thinking
240	ISRN Radiology	International Journal of Medical Reviews
241	ISRN Rehabilitation	International Journal of Modern Anthropology
242	ISRN Rheumatology	International Journal of Multicultural and Multireligious Understanding
243	ISRN Stroke	International Journal of Public Information Systems
244	ISRN Toxicology	International Journal of Research in Business and Social Science
245	ISRN Transplantation	International Journal of Sport Management, Recreation and Tourism
246	ISRN Urology	International Journal on Working Conditions
247	ISRN Vascular Medicine	International Review of Social Research
248	ISRN Virology	International Review of Social Sciences and Humanities
249	Italian Journal of Medicine	Internet Journal of Criminology
250	JMIR Research Protocols	Intersections
251	JMM Case Reports	Investigaciones Regionales

continued ...

... continued

No.	Medicine	Social sciences
252	Journal of Advanced Pharmaceutical Technology and Research	Iranian Journal of Management Studies
253	Journal of Anaesthesiology Clinical Pharmacology	Irish Journal of Applied Social Studies
254	Journal of Applied Biotechnology Reports	İşletme Araştırmaları Dergisi
255	Journal of Arthropod-Borne Diseases	ISRN Addiction
256	Journal of Biomedical Science	ISRN Economics
257	Journal of BioScience and Biotechnology	Issues in Informing Science and Information Technology
258	Journal of Cancer Epidemiology	Issues in Science and Technology Librarianship
259	Journal of Cancer Research and Therapy	Istanbul Gelisim University Journal of Social Sciences
260	Journal of Cancer Research and Therapeutics	Italian Journal of Sociology of Education
261	Journal of Carcinogenesis	Janus.net
262	Journal of Cellular and Molecular Medicine	Journal for Communication and Culture
263	Journal of Chiropractic Education	Journal for Deradicalization
264	Journal of Clinical and Analytical Medicine	Journal Of Accounting, Finance and Auditing Studies
265	Journal of Clinical and Diagnostic Research	Journal of Advanced Ceramics
266	Journal of Clinical and Experimental Investigations	Journal of Advocacy, Research and Education
267	Journal of Clinical Medicine	Journal of Agriculture and Environment for International Development
268	Journal of Comorbidity	Journal of Airline and Airport Management

continued ...

... continued

No.	Medicine	Social sciences
269	Journal of Computational Surgery	Journal of Artificial Societies and Social Simulation
270	Journal of Conservative Dentistry	Journal of Arts and Social Sciences
271	Journal of Contemporary Brachytherapy	Journal of ASEAN Studies
272	Journal of Craniovertebral Junction and Spine	Journal of Balkan Libraries Union
273	Journal of Cutaneous and Aesthetic Surgery	Journal of Blindness Innovation and Research
274	Journal of Dentistry	Journal of Contemporary European Research
275	Journal of Dentistry Indonesia	Journal of Copyright in Education and Librarianship
276	Journal of Diabetology Official Journal of Diabetes in Asia Study Group	Journal of Current Southeast Asian Affairs
277	Journal of Education, Health and Sport	Journal of Data Mining and Digital Humanities
278	Journal of Educational Evaluation for Health Professions	Journal of Economics and Behavioral Studies
279	Journal of Emergencies, Trauma and Shock	Journal of Environmental Hydrology
280	Journal of Enam Medical College	Journal of Forensic Science and Criminology
281	Journal of Experimental Orthopaedics	Journal of Global Analysis
282	Journal of Fitness Research	Journal of Government and Politics
283	Journal of Food and Pharmaceutical Sciences	Journal of Human Growth and Development
284	Journal of Forensic Dental Sciences	Journal of Human Security
285	Journal of Global Health	Journal of Humanistics and Social Sciences

continued ...

... continued

No.	Medicine	Social sciences
286	Journal of Global Infectious Diseases	Journal of Industrial Engineering and Management
287	Journal of Hormones	Journal of Information Literacy
288	Journal of Indian Association of Pediatric Surgeons	Journal of Innovations and Sustainability
289	Journal of Interdisciplinary Histopathology	Journal of International Trade, Logistics and Law
290	Journal of International Child Neurology Association	Journal of Islamic Architecture
291	Journal of International Translational Medicine	Journal of Islamic Banking and Finance
292	Journal of Kermanshah University of Medical Sciences	Journal of Knowledge Management, Economics and Information Technology
293	Journal of Krishna Institute of Medical Sciences University	Journal of Librarianship and Scholarly Communication
294	Journal of Laboratory Physicians	Journal of Library and Information Studies
295	Journal of Liaquat University of Medical and Health Sciences	Journal of Mediterranean Knowledge
296	Journal of Medical and Allied Sciences	Journal of Methods and Measurement in the Social Sciences
297	Journal of Medical and Surgical Research	Journal of Organization Design
298	Journal of Medical Case Reports	Journal of Physical Education and Sport
299	Journal of Medical Physics	Journal of Political Ecology
300	Journal of Medical Sciences	Journal of Political Studies
301	Journal of Military and Veterans' Health	Journal of Politics in Latin America

continued ...

... continued

No.	Medicine	Social sciences
302	Journal of Minimal Access Surgery	Journal of Quality and Reliability Engineering
303	Journal of Minimally Invasive Surgical Sciences	Journal of Science and Technology of the Arts
304	Journal of Negative Results in Biomedicine	Journal of Service Science
305	Journal of Neonatal Surgery	Journal of Smart Economic Growth
306	Journal of Neurosciences in Rural Practice	Journal of Social and Development Sciences
307	Journal of Nutritional Science	Journal of Social Research and Policy
308	Journal of Occupational Therapy	Journal of Sustainable Development of Energy, Water and Environment Systems
309	Journal of Oral and Maxillofacial Pathology	Journal of Systems Integration
310	Journal of Oral Diseases	Journal of the International AIDS Society
311	Journal of Oral Research	Journal of Transnational American Studies
312	Journal of Osseointegration	Journal of Transport and Land Use
313	Journal of Osteoporosis	Journal of World-Systems Research
314	Journal of Pakistan Medical Students	Jurnal Dinamika Manajemen
315	Journal of Patient-Centered Research and Reviews	Khazar Journal of Humanities and Social Sciences
316	Journal of Pediatric and Neonatal Individualized Medicine	Kōtuitui New Zealand Journal of Social Sciences Online
317	Journal of Pediatric Emergency and Intensive Care Medicine	Kultura
318	Journal of Pediatric Neurosciences	Language and Literacy
319	Journal of Pediatric Research	Law, Crime and History

continued ...

. . . continued

No.	Medicine	Social sciences
320	Journal of Personalized Medicine	Lithuanian Annual Strategic Review
321	Journal of Pharmaceutical Health Care and Sciences	Management Journal of Contemporary Management Issues
322	Journal of Pharmaceutical Research and Health Care	Marketing of Scientific and Research Organisations
323	Journal of Pharmacology and Pharmacotherapeutics	Medieval Worlds
324	Journal of Pharmacopuncture	methaodos.revista de ciencias socials
325	Journal of Physiotherapy and Sports Medicine	Mobile Culture Studies. The Journal
326	Journal of Pioneering Medical Sciences	Momentum Quarterly
327	Journal of Postgraduate Medicine	Mountain Research and Development
328	Journal of Preventive Medicine and Public Health	Moussons
329	Journal of Research in Medical Sciences	Multidisciplinary Journal for Education, Social and Technological Sciences
330	Journal of Rural and Tropical Public Health	Neo-Victorian Studies
331	Journal of Special Education and Rehabilitation	Nómadas
332	Journal of Stem Cells and Regenerative Medicine	Nordic Journal of Social Research
333	Journal of the Indian Society of Pedodontics and Preventive Dentistry	OBETS. Revista de Ciencias Sociales
334	Journal of the Indian Society of Periodontology	Œconomia
335	Journal of the International AIDS Society	Oeconomia Copernicana
336	Journal of the Medical Library Association	On Our Terms

continued . . .

... continued

No.	Medicine	Social sciences
337	Journal of the Medical Sciences	Outskirts feminisms along the edge
338	Journal of the Royal College of Physicians of Edinburgh	Pakistan Journal of Commerce and Social Sciences
339	Journal of the Scientific Society	Palgrave Communications
340	Journal of Traditional and Complementary Medicine	Papers Revista de Sociologia
341	Journal of Translational Medicine	Peregrinations
342	Journal of Vector Borne Diseases	Persona y Bioética
343	Jurnal Kesehatan Reproduksi	Perspectives of Innovations, Economics and Business
344	Jurnal Sains Kesihatan Malaysia	Politics in Central Europe
345	KEMAS Journal of Public Health	Postmodern Openings
346	Kerala Heart Journal	Pragmatic Case Studies in Psychotherapy
347	Kesmas Jurnal Kesehatan Masyarakat Nasional	Proceedings of Rijeka Faculty of Economics
348	Koşuyolu Heart Journal	PSL Quarterly Review
349	Libyan Journal of Medicine	Psychological Test and Assessment Modeling
350	Lung India	QJB Querelles
351	Makara Journal of Health Research	Rationality, Markets and Morals
352	Maternal Health, Neonatology and Perinatology	Red Feather Journal
353	Meandros Medical and Dental Journal	Research and Politics
354	Medical Express	Review of Economics and Finance
355	Medical Journal of Dr. D.Y. Patil University	Revija za socijalnu politiku
356	Medical Sciences	Revista Brasileira de Política Internacional
357	Medicina	Revista de Administração Mackenzie

continued ...

. . . continued

No.	Medicine	Social sciences
358	Medicine	Revista Economică
359	Medicines	Revue LISA
360	Medicinski Glasnik	RIHA Journal
361	Medicinski Glasnik Specijalne Bolnice za Bolesti	RSF The Russell Sage Foundation Journal of the Social Sciences
362	Menopause Review	Rupkatha Journal on Interdisciplinary Studies in Humanities
363	Mens Sana Monographs	SAGE Open
364	Mental Illness	Science and Philosophy
365	Middle East African Journal of Ophthalmology	Secrecy and Society
366	Molecular and Cellular Epilepsy	Securitologia
367	Molecular Imaging and Radionuclide Therapy	Selçuk Üniversitesi Edebiyat Fakültesi Dergisi
368	Motricidade	SENTENTIA. European Journal of Humanities and Social Sciences
369	Nanobiomedicine	SERIEs
370	Neurological Journal of South East Asia	Sexual Offender Treatment
371	Neurology and Therapy	Shodh Sanchayan
372	Neurology India	SHS Web of Conferences
373	Neurology International	Slovak Journal of Political Sciences
374	Neuropsychiatric Disease and Treatment	Social Affairs
375	Neurovascular Imaging	Social Science Diliman
376	New Medicine	Social Sciences
377	Nigerian Journal of Surgery	Social Space
378	Nigerian Medical Journal	Social Transformations in Contemporary Society
379	Noise and Health	Socialist Studies

continued . . .

... continued

No.	Medicine	Social sciences
380	North American Journal of Medical Sciences	Socioeconomica
381	npj Genomic Medicine	Socio-Legal Review
382	NSW Public Health Bulletin	Sociology and Criminology
383	Nuclear Medicine Review	SoundEffects
384	Nutrition and Dietary Supplements	South African Journal of Science
385	Odontoestomatología	South Asia Multidisciplinary Academic Journal
386	Oman Journal of Ophthalmology	South Asian Studies
387	Oman Medical Journal	South East Asian Journal of Management
388	Oncology in Clinical Practice	South-East European Journal of Political Science
389	OncoTargets and Therapy	Sprawy Narodowościowe
390	Online Journal of Health and Allied Sciences	SQS Journal of Queer Studies in Finland
391	Online Journal of Nursing Informatics	Studia Humanistyczne
392	Open Access Journal of Clinical Trials	Studia Universitatis Moldaviae Stiinte Sociale
393	Open Access Journal of Urology	Studies in History and Theory of Architecture
394	Open Access Macedonian Journal of Medical Sciences	Studies of Transition States and Societies
395	Open Health Data	Studies on Asia
396	Open Journal of Bioresources	Sustainability Science, Practice and Policy
397	Open Medicine	Symposion
398	Optometry and Visual Performance	Tate Papers
399	Orphanet Journal of Rare Diseases	Technology Audit and Production Reserves

continued ...

... continued

No.	Medicine	Social sciences
400	Otolaryngology Online Journal	Terengganu International Management and Business Journal
401	Oxford Medical Case Reports	The Arbutus Review
402	Paediatrica Indonesiana	The International Journal of Evidence Based Coaching and Mentoring
403	Pakistan Armed Forces Medical Journal	The Journal of Chinese Sociology
404	Pakistan Journal of Medical Sciences	The Journal of Philosophical Economics
405	Pakistan Journal of Ophthalmology	The Romanian Economic Journal
406	Pakistan Journal of Pharmaceutical Research	The Scottish Journal of Performance
407	Panacea Journal of Medical Sciences	The USV Annals of Economics and Public Administration
408	Pathogens	Theoretical and Applied Economics
409	Patient Related Outcome Measures	Theory, Methodology, Practice
410	Pediatric Health, Medicine and Therapeutics	Torun International Studies
411	Pediatric Rheumatology	Trans-Asia Photography Review
412	PeerJ	Transcultural Studies
413	People Living with And Inspired by Diabetes	TransNav
414	Perspectives in Clinical Research	Üniversitepark Bülten
415	Perspectives In Medical Research	URBS Revista de Estudios Urbanos y Ciencias Sociales
416	Pharmacognosy Research	Vestnik The Journal of Russian and Asian Studies
417	Physical Education of Students	WACANA
418	PLoS Medicine	
419	PLoS ONE	

continued ...

. . . continued

No.	Medicine	Social sciences
420	Polish Journal of Pathology	
421	Polish Journal of Surgery	
422	Polish Journal of Thoracic and Cardiovascular Surgery	
423	Porcine Health Management	
424	Pragmatic Case Studies in Psychotherapy	
425	Preventive medicine reports	
426	Progress in Health Sciences	
427	Psoriasis Targets and Therapy	
428	Psychiatria i Psychologia Kliniczna	
429	Public Health of Indonesia	
430	Rambam Maimonides Medical Journal	
431	Rare Tumors	
432	Razavi International Journal of Medicine	
433	Recent Advances in Biology and Medicine	
434	Regenerative Medicine Research	
435	Reports in Medical Imaging	
436	Research and Reports in Endocrine Disorders	
437	Research and Reports in Urology	
438	Research in Pharmaceutical Sciences	
439	Research Involvement and Engagement	
440	Reumatismo	
441	Reumatologia	
442	Revista Andaluza de Medicina del Deporte	
443	Revista Argentina de Cardiología	

continued . . .

. . . continued

No.	Medicine	Social sciences
444	Revista Brasileira de Ortopedia	
445	Revista de Enfermagem Referência	
446	Revista de Odontologia da UNESP	
447	Revista de Osteoporosis y Metabolismo Mineral	
448	Revista de Patologia Tropical	
449	Revista de Pesquisa Cuidado é Fundamental Online	
450	Revista de Toxicología	
451	Revista Dor	
452	Revista Eletrônica de Enfermagem	
453	Revista Española de Sanidad Penitenciaria	
454	Revista Médica Clínica Las Condes	
455	RIASE	
456	RMD Open	
457	RNA & DISEASE	
458	Romanian Journal of Laboratory Medicine	
459	Romanian Journal of Military Medicine	
460	Rural and Remote Health	
461	São Paulo Medical Journal	
462	Saudi Dental Journal	
463	Saudi Journal of Gastroenterology	
464	Saudi Journal of Kidney Diseases and Transplantation	
465	Science Postprint	
466	Scientific Reports	
467	Senses and Sciences	

continued . . .

. . . continued

No.	Medicine	Social sciences
468	Serbian Journal of Experimental and Clinical Research	
469	Sexual Medicine	
470	Signal Transduction and Targeted Therapy	
471	South Afican Journal of Child Health	
472	South African Journal of Bioethics and Law	
473	South African Journal of Obstetrics and Gynaecology	
474	South African Medical Journal	
475	South Asian Journal of Cancer	
476	Southern Afican Journal of Critical Care	
477	Sports Medicine – Open	
478	Sports Medicine, Arthroscopy, Rehabilitation, Therapy and Technology	
479	Sri Lankan Journal of Anaesthesiology	
480	Srpski Arhiv za Celokupno Lekarstvo	
481	Strategies in Trauma and Limb Reconstruction	
482	Studia Medyczne	
483	Studia Universitatis Vasile Goldis	
484	Substance Abuse and Rehabilitation	
485	Sultan Qaboos University Medical Journal	
486	Surgical Case Reports	
487	Surgical Neurology International	
488	Swiss Medical Weekly	
489	Systematic Reviews	

continued . . .

... continued

No.	Medicine	Social sciences
490	TAF Preventive Medicine Bulletin	
491	Texto and Contexto Enfermagem	
492	The European Journal of Psychiatry	
493	The Indonesian Journal of Gastroenterology, Hepatology and Digestive Endoscopy	
494	The Journal of Faculty of Medicine in Nis	
495	The Turkish Nephrology, Dialysis and Transplantation Journal	
496	Theranostics	
497	Therapeutic Targets for Neurological Diseases	
498	Therapeutics and Clinical Risk Management	
499	Toxins	
500	Traditional Medicine Research	
501	Transplant Research and Risk Management	
502	Tropical Medicine and Infectious Disease	
503	Tuberculosis Research and Treatment	
504	Türk Hijyen ve Deneysel Biyoloji Dergisi	
505	Türk Oftalmoloji Dergisi	
506	Türk Osteoporoz Dergisi	
507	Turk Dermatoloji Dergisi	
508	Turkderm	
509	Turkish Journal of Clinics and Laboratory	

continued ...

. . . continued

No.	Medicine	Social sciences
510	Turkish Journal of Plastic Surgery	
511	Turkiye Klinikleri Journal of Biostatistics	
512	Üroonkoloji Bülteni	
513	Ultrasound International Open	
514	University of Ottawa Journal of Medicine	
515	Urology Annals	
516	Vaccines	
517	Vascular Cell	
518	Veterinary Science and Medicine Journal	
519	Videosurgery and Other Miniinvasive Techniques	
520	Viral Hepatitis Journal	
521	Western Journal of Emergency Medicine	
522	Western Pacific Surveillance and Response	
523	Women's Midlife Health	
524	World Journal of Clinical Cases	
525	World Journal of Clinical Pediatrics	
526	World Journal of Nephrology and Urology	
527	World Journal of Traditional Chinese Medicine	
528	Yemeni Journal for Medical Sciences	
529	Zahedan Journal of Researche in Medical Sciences	
530	Zdravniški Vestnik	

Appendix H

Code

Following code snippets were used to implement the new recommender system using Lucene API. The given example code shows computations for the BM25 algorithm.

```
/******** Libraries used ***********/

import java.io.File;
import java.io.FileReader;
import java.io.IOException;
import java.io.Reader;
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.ResultSet;
import java.sql.ResultSetMetaData;
import java.sql.Statement;
import java.util.Scanner;
import java.util.Arrays;

import org.apache.lucene.analysis.Analyzer;
import org.apache.lucene.analysis.WhitespaceAnalyzer;
```

```
import org.apache.lucene. analysis .snowball.SnowballAnalyzer;
import org.apache.lucene.document.Document;
import org.apache.lucene.document.Field;
import org.apache.lucene.index.CorruptIndexException;
import org.apache.lucene.index.IndexReader;
import org.apache.lucene.index.IndexWriter;
import org.apache.lucene.index.TermFreqVector;
import org.apache.lucene.queryParser.ParseException;
import org.apache.lucene.queryParser.QueryParser;
import org.apache.lucene.search.IndexSearcher;
import org.apache.lucene.search.Query;
import org.apache.lucene.store. Directory;
import org.apache.lucene.store.FSDirectory;
import org.apache.lucene.store.LockObtainFailedException;
import org.apache.commons.lang.ArrayUtils;

public class BestMat1 {

  /********** Index directories for input abstract and corpus **********/

public static final String FILES_TO_INDEX_DIRECTORY1 = "filesToIndex1";
public static final String INDEX_DIRECTORY1 = "indexDirectory1";
public static final String FIELD_PATH1 = "path";
public static final String FIELD_CONTENTS1 = "contents";
public static final String FILES_TO_INDEX_DIRECTORY2 = "filesToIndex2";
public static final String INDEX_DIRECTORY2 = "indexDirectory2";
public static final String FIELD_PATH2 = "path";
public static final String FIELD_CONTENTS2 = "contents";

/***********Create index*************/
```

```
public static void createIndex() throws CorruptIndexException,
    LockObtainFailedException, IOException {

  final String[] NEW_STOP_WORDS = {.....}"stop words"{.....};
    SnowballAnalyzer analyzer = new SnowballAnalyzer("English",
        NEW_STOP_WORDS );
    Directory directory = FSDirectory.getDirectory(INDEX_DIRECTORY1);
    IndexWriter w = new IndexWriter(INDEX_DIRECTORY1, analyzer,
        true,IndexWriter.MaxFieldLength.UNLIMITED);
    File dir = new File(FILES_TO_INDEX_DIRECTORY1);
    File[] files = dir.listFiles();

  for (File file : files) {
      Document doc = new Document();
      String textb = "";
            doc.add(new Field("contents", text, Field.Store.YES,
                Field.Index.UN_TOKENIZED,Field.TermVector.YES));
      Reader reader = new FileReader(file);
      doc.add(new Field(FIELD_CONTENTS1, reader));
            w.addDocument(doc);
    }
    w.optimize();
    w.close();
  }

/**********Search index**********/
public static void searchIndex() throws IOException, ParseException {

Directory directory = FSDirectory.getDirectory(INDEX_DIRECTORY1);
IndexReader ir1 = IndexReader.open(directory);
```

```
TermFreqVector[] tfv = ir1.getTermFreqVectors(0);
String [] terms = tfv[0].getTerms();
int [] freqs = tfv[0].getTermFrequencies();
int que_ter;
int que_len;
que_len = freqs.length;
int corpus = 4170; //Corpus size
int doc;
double BM25_score[] = new double[corpus];
double tf1_doc[][] = new double[corpus][que_len];
double tf_doc[][] = new double[corpus][que_len];
double score[][] = new double[corpus][que_len];
BestMat2.totalfre(); //Calling totalfre variable from BestMat2 class
Directory directory2 = FSDirectory.getDirectory(INDEX_DIRECTORY2);
IndexReader ir2 = IndexReader.open(directory2);
IndexSearcher indexSearcher2 = new IndexSearcher(ir2);
Analyzer analyzer2 = new WhitespaceAnalyzer();
QueryParser queryParser2 = new QueryParser(FIELD_CONTENTS2, analyzer2);
int hits_queter[] = new int[que_len]; //Number of hits for query document terms
    in the corpus
double idf1_doc[] = new double[que_len]; //idf values of the query document
    terms in the corpus

  for(que_ter=1;que_ter<que_len;que_ter++){
      Query query = queryParser2.parse(terms[que_ter]);
          Hits hits = indexSearcher2.search(query);
            hits_queter[que_ter] = hits.length();

            /**********Computing idf component**********/
```

```
                idf1_doc[que_ter] =
                Math.log(1.0+(((double)corpus-(double)hits_queter[que_ter]+0.5)
                /((double)hits_queter[que_ter]+0.5)));
            }
/********** Computing BM25 score ***********/
for (doc=0;doc<corpus;doc++){

  BM25_score[doc] = 0.0;

    for(que_ter=1;que_ter<que_len;que_ter++){ //Check for all query terms in
        the corpus
      if( ArrayUtils.contains(ir2.getTermFreqVectors(doc)[0].getTerms(),
                terms[que_ter])){
          tf1_doc[doc][que_ter]
                =ir2.getTermFreqVectors(doc)[0].getTermFrequencies()[que_ter];
          tf_doc[doc][que_ter]=tf1_doc[doc][que_ter];
      }
       else{
          tf_doc[doc][que_ter]=0;
      }
    score[doc][que_ter]=(2.2*tf_doc[doc][que_ter]*idf1_doc[que_ter])
    /(0.3+(0.9*BestMat2.totfre[doc]/BestMat2.avg_dlen)+tf_doc[doc][que_ter]);
    BM25_score[doc]=BM25_score[doc]+score[doc][que_ter];

   }
}

/***Connecting with Excel database to access journal metadata***/
try {
     Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");
```

```
    }
    catch (Exception e) {
     System.err. println (e);
    }
Connection conn=null;
Statement stmt=null;
String sql="";
ResultSet rs=null;
double [][] colVal1=new double[17][10];   //17 stands for (number of factors+2),
    10 stands for topmost journals
double[] Similarity =new double[10];
double[] TotSimilarity =new double[10];
double[] FinalTotSimilarity =new double[10];
double[] UserInput=new double[15]; //15=number of factors
double[] TotWeight=new double[15];
double[] RangeData={113,53,4,6.296,2.216,2.74,100,100,1059,24}; //Range of each
    factor value in the database, except for binary valued factors
double[] Weight ={2.59,3.71,3.84,3.77,3.94,3.10,3.38,3.16,
     2.80,2.86,3.13,3.92,3.14,4.59,3.34};
    //Weights for 15 factors
Scanner factorreader =new Scanner(System.in);
 for (int fac = 0; fac <15; fac++){

    System.out. print ("Enter author's choice for factor "+(fac+1)+":");
  UserInput[fac]= factorreader .nextDouble();
  }

factorreader . close ();
String [] JournalName=new String[10];
```

```
for (int index1=0; index1<10; index1++){


    try {
        conn = DriverManager.getConnection("jdbc:odbc:excel","","");
      stmt=conn.createStatement();
        sql="select * from [Sheet1$] where
            JOURNAL_TITLE='"+file_new[position[journals-(index1+1)]]+"'";
    rs=stmt.executeQuery(sql);
    ResultSetMetaData rsmd = rs.getMetaData();
    int numberOfColumns = rsmd.getColumnCount();


    while (rs.next()) {
            for (int column = 2; column <= numberOfColumns; column++) {
                colVal1[column][index1]=rs.getDouble(column);
        }
        System.out.println ("");
     }
}
catch (Exception e){
            System.err.println (e);
}
 finally {
     try{
                rs.close();
                stmt.close();
                conn.close();
                rs=null;
                stmt=null;
                conn=null;
     }
```

```
    catch(Exception e){}
}

  /*************** Computing Gower's measure *************/

 TotSimilarity [index1]=0.0;
TotWeight[0]=0.0;
  int inp;
  for(inp=0; inp<10; inp++){ //10=number of factors without binary values
  if(UserInput[inp]>=0){
            Similarity [index1]=1-((Math.abs(UserInput[inp]
                  -colVal1[inp+2][index1]))/RangeData[inp]);
   }
    else{
        Similarity [index1]=0.0;
       Weight[inp]=0.0;
   }
    TotSimilarity [index1]= TotSimilarity [index1]+Weight[inp]* Similarity [index1];
   TotWeight[index1]=TotWeight[index1]+Weight[inp];
  }

for(inp=10;inp<15;inp++){ //10 to 15 numbers of factors with binary values
     if((UserInput[inp]==1)&&(colVal1[inp+2][index1]==1)){
        Similarity [index1]=1.0;
  }
       else if(UserInput[inp]==-1){// -1 if the author does not consider a
            particular factor
        Similarity [index1]=0.0;
       Weight[inp]=0.0;
  }
```

```
      else{
            Similarity [index1]=0.0;
      }
            TotSimilarity [index1]= TotSimilarity [index1]+Weight[inp]* Similarity [index1];
      TotWeight[index1]=TotWeight[index1]+Weight[inp];
      }

     FinalTotSimilarity [index1]= TotSimilarity [index1]/TotWeight[index1];
      System.out. println (" Similarity ["+index1+"]="+FinalTotSimilarity[index1]+"="
      +file_new[position[ journals -(index1+1)]]);
    JournalName[index1]=file_new[position[ journals -(index1+1)]];
      }
   }
}
```